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

Project Deliverable 2

All project questions should have been reviewed, answered, and submitted in a written document outlining findings as a PR. You will also be asked to submit the associated data and a README explaining what each label/feature in your dataset represents.

****

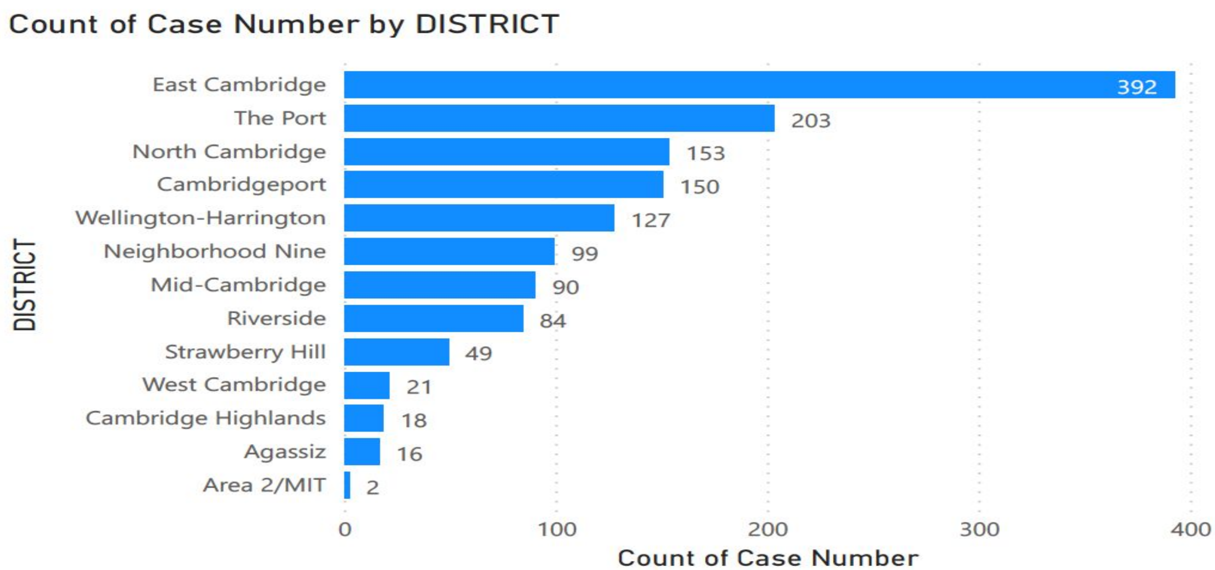
The following strategic questions have been reviewed, answered, and explained. Results can be reproduced using the final dataset. Data was retrieved from the databases provided at the beginning of this project and the additional web databases cited at the end of this deliverable. The final dataset shows defendant cases (rows) and associated features (columns).

Strategic Questions:

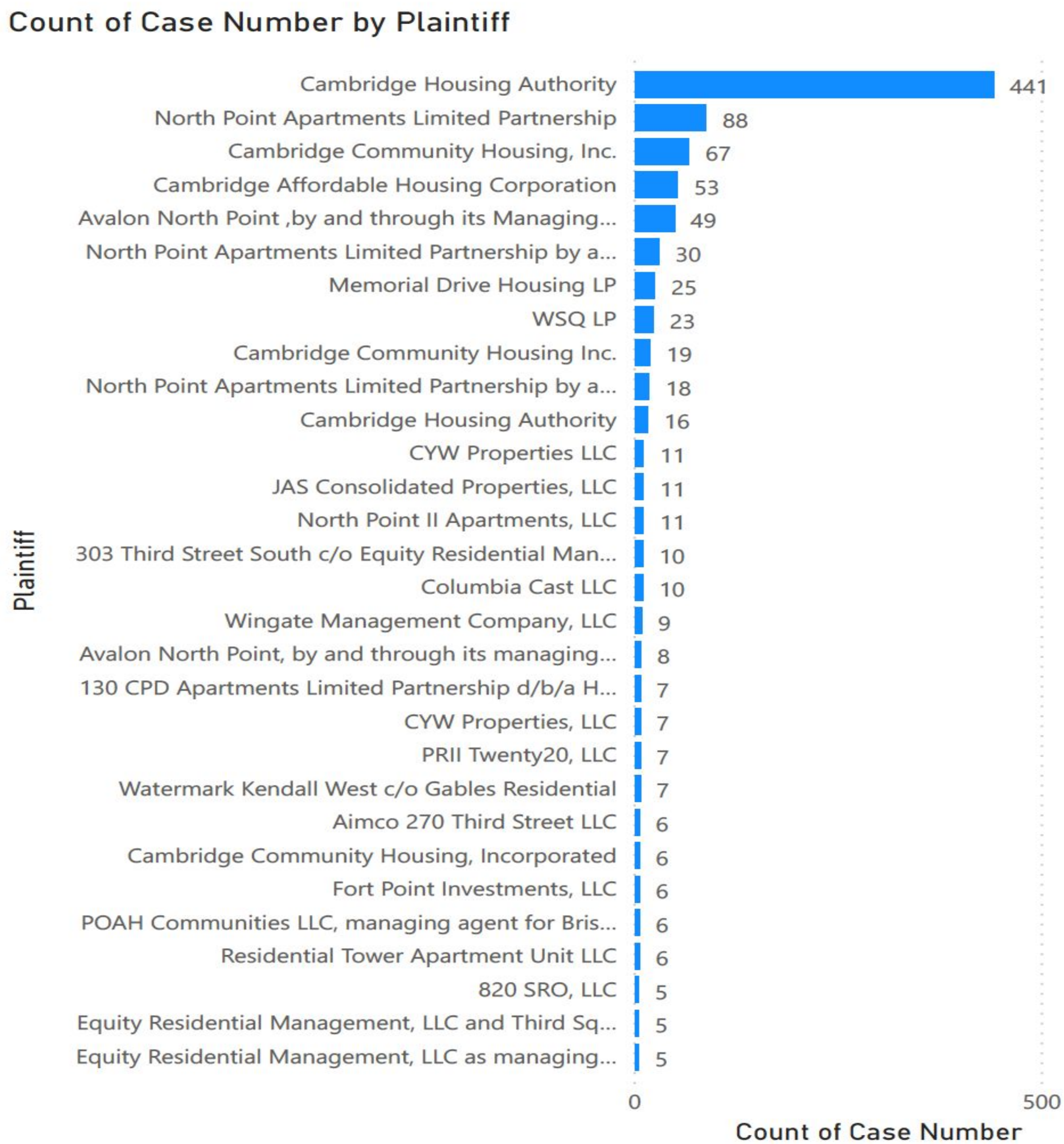
1. Are there local demographic, physical housing conditions or spatial characteristics that make units more likely to be subject to eviction?

   Based on our analysis, the eviction cases happen mostly for people who live in the district one (East Cambridge) and for people who rent from Cambridge public housing authority.

   From the eviction data from 2017 to the beginning of 2020 there are 1404 cases of eviction in the city of cambridge. Around 70% of total eviction cases happened in just 4 districts-- East Cambridge, The Port, North Cambridge, and Cambridge Port-- with the most cases occurring in district East Cambridge as depicted in the figure below.

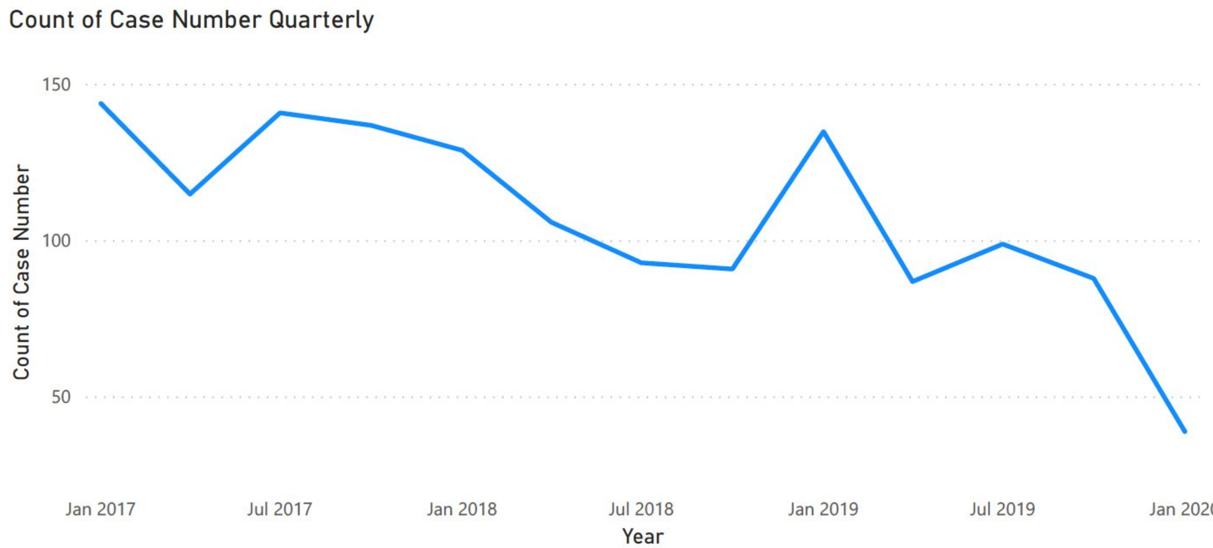


Most of the eviction cases occurred in the property owned by Cambridge Housing Authority.



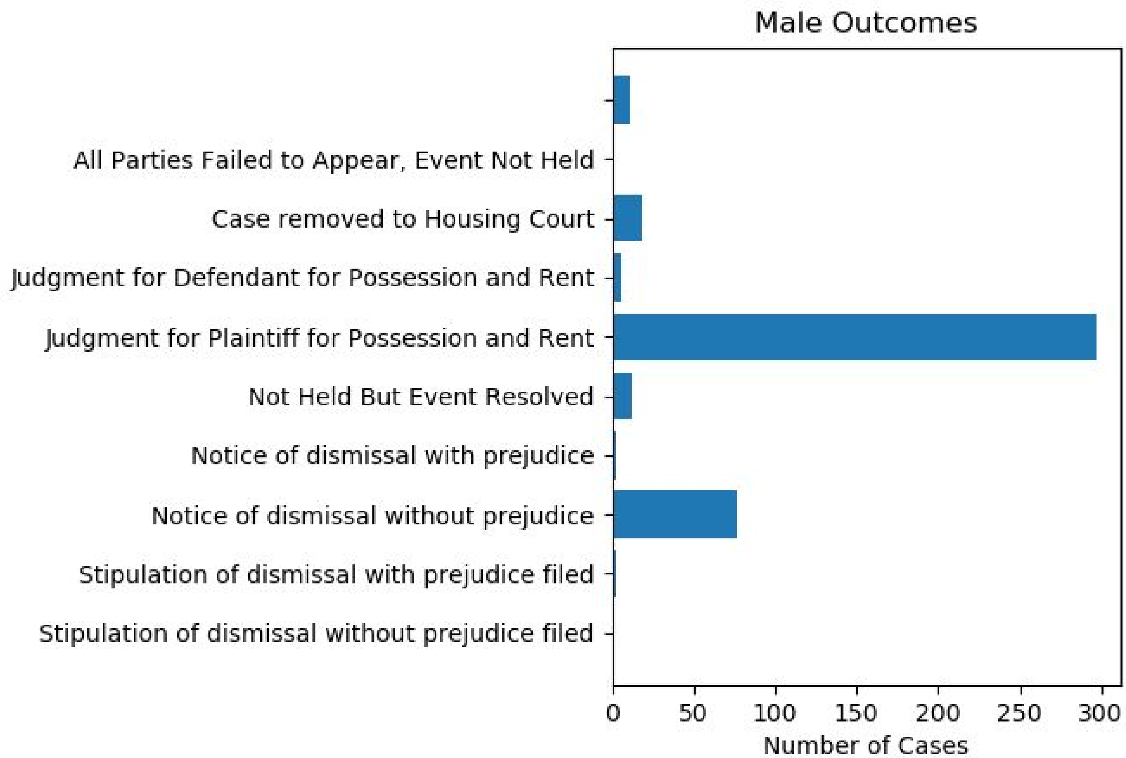
The number of eviction cases reduced over time. For 2020, the data is not yet complete.

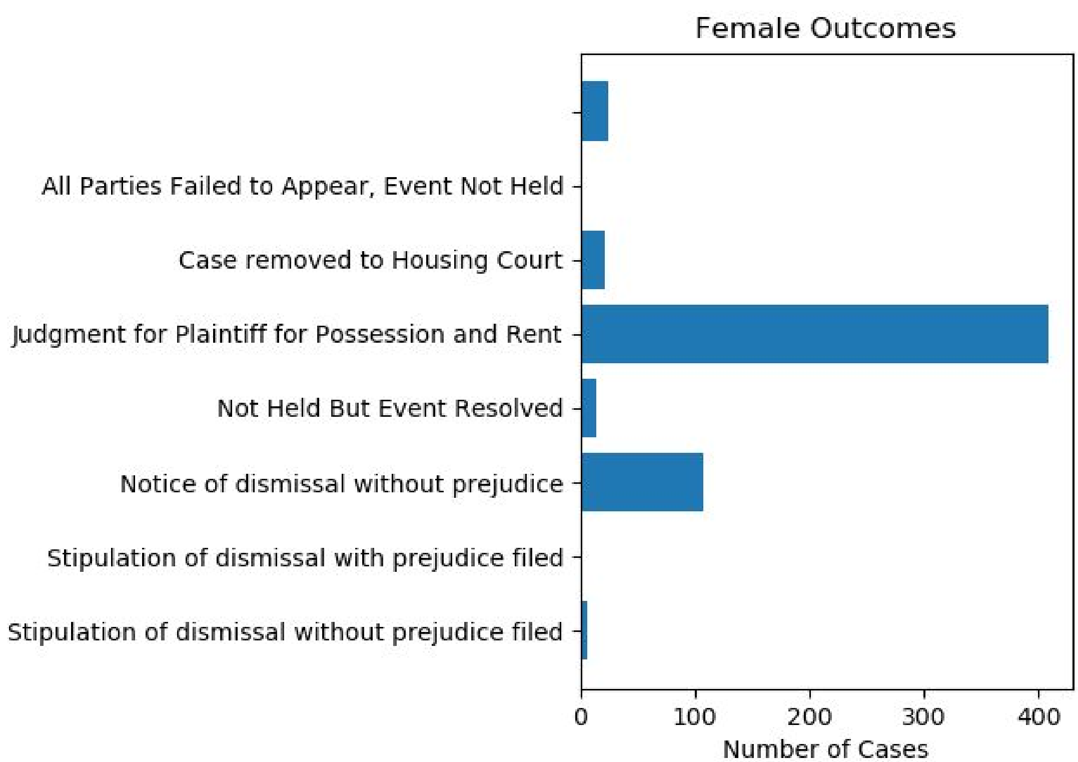
Additionally, the eviction cases can be analyzed by feature in the dataset and clustered.



For Question 1, we decided to examine how cases could be clustered depending on specific features. We were also able to predict gender and race using the “gender-guesser” and “ethnicolr” packages, respectively. Gender was determined for each name in the database. However, gender-guesser can be used to classify names from certain regions better than it can classify names from other regions. We are investigating using an API to supplement our gender classification feature. An initial attempt at using the gender feature for clustering is shown below. Still, even without using an API, most of the names have been classified as male or female.

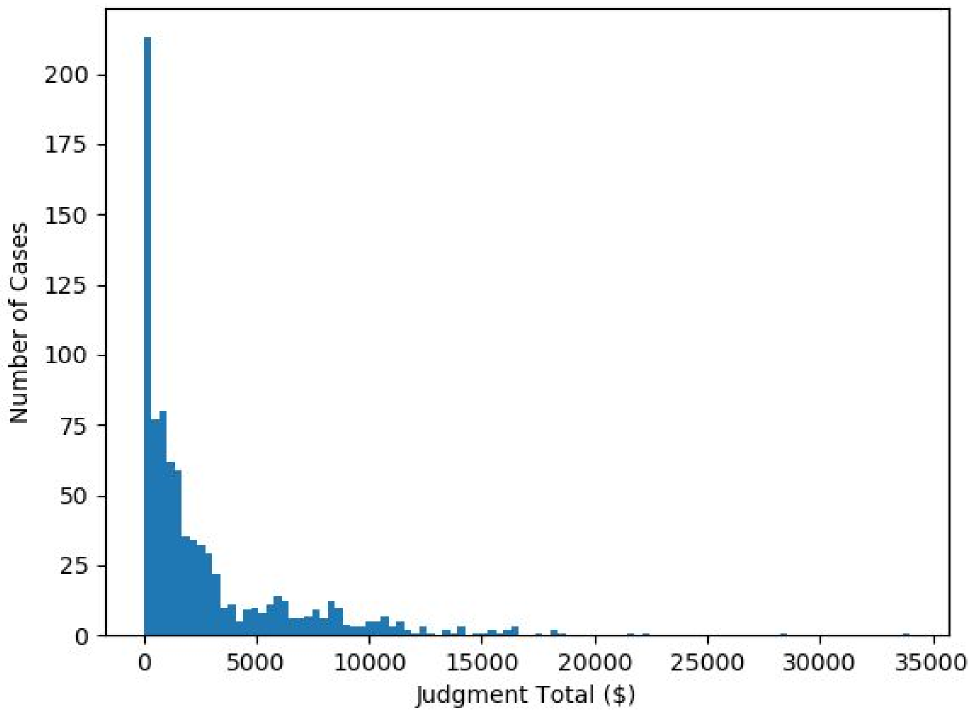
There are not any major differences between male and female outcomes.

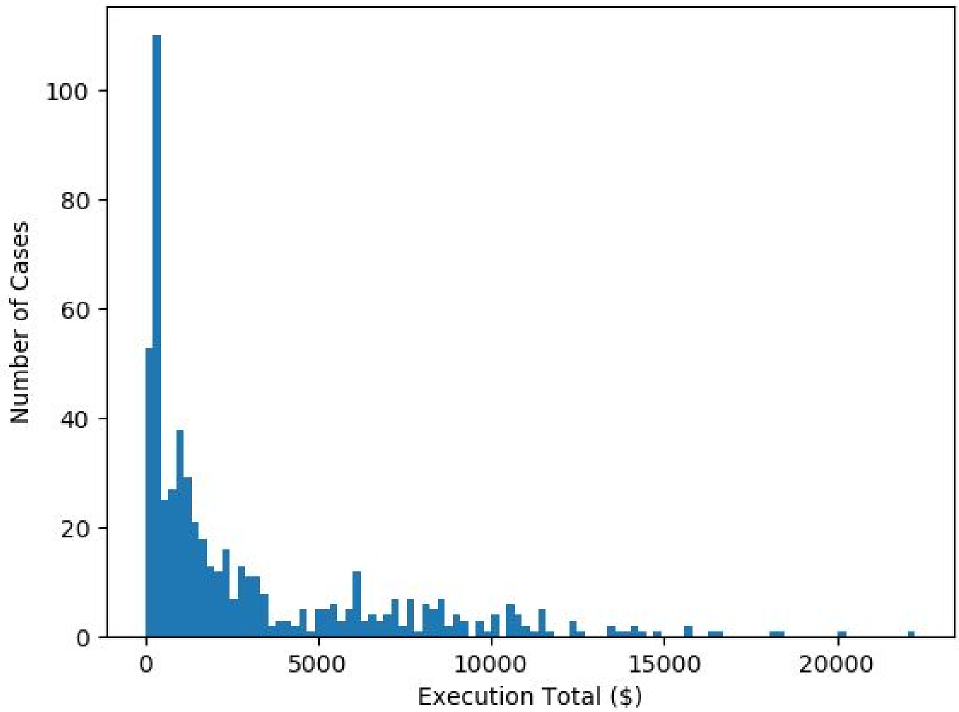

Male Outcomes
All Parties Failed to Appear, Event Not Held
Case removed to Housing Court
Judgment for Defendant for Possession and Rent
Judgment for Plaintiff for Possession and Rent
Not Held But Event Resolved
Notice of dismissal with prejudice
Notice of dismissal without prejudice
Stipulation of dismissal with prejudice filed
Stipulation of dismissal without prejudice filed
0
50
100
150
200
250
300
Number of Cases


Female Outcomes
All Parties Failed to Appear, Event Not Held
Case removed to Housing Court
Judgment for Plaintiff for Possession and Rent
Not Held But Event Resolved
Notice of dismissal without prejudice
Stipulation of dismissal with prejudice filed
Stipulation of dismissal without prejudice filed
0
100
200
300
400
Number of Cases

Additionally, ethnicolr relies heavily on Tensorflow, a large package that relies on many other packages. At the time of the submission of this deliverable, our team has a plan to implement the ethnicolr package to determine, with reasonable confidence, the race and ethnicity of each defendant, but difficulties in installing the package on a Windows PC have hindered this progress. We hope to remedy this by the time of final submission so that we can examine defendant cases with another feature.

Histograms for the judgment totals and execution totals are shown below.





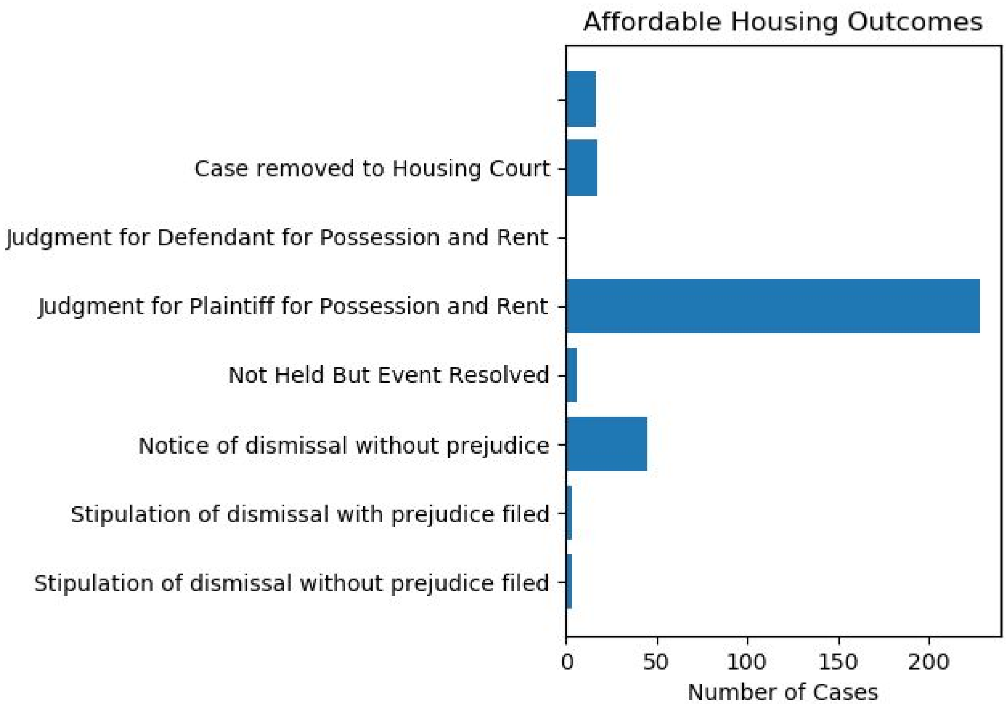
2. Are there distinct differences between housing that occur in the low-rise and small scale housing stock (developments under 25 units) compared to the newer, larger buildings?

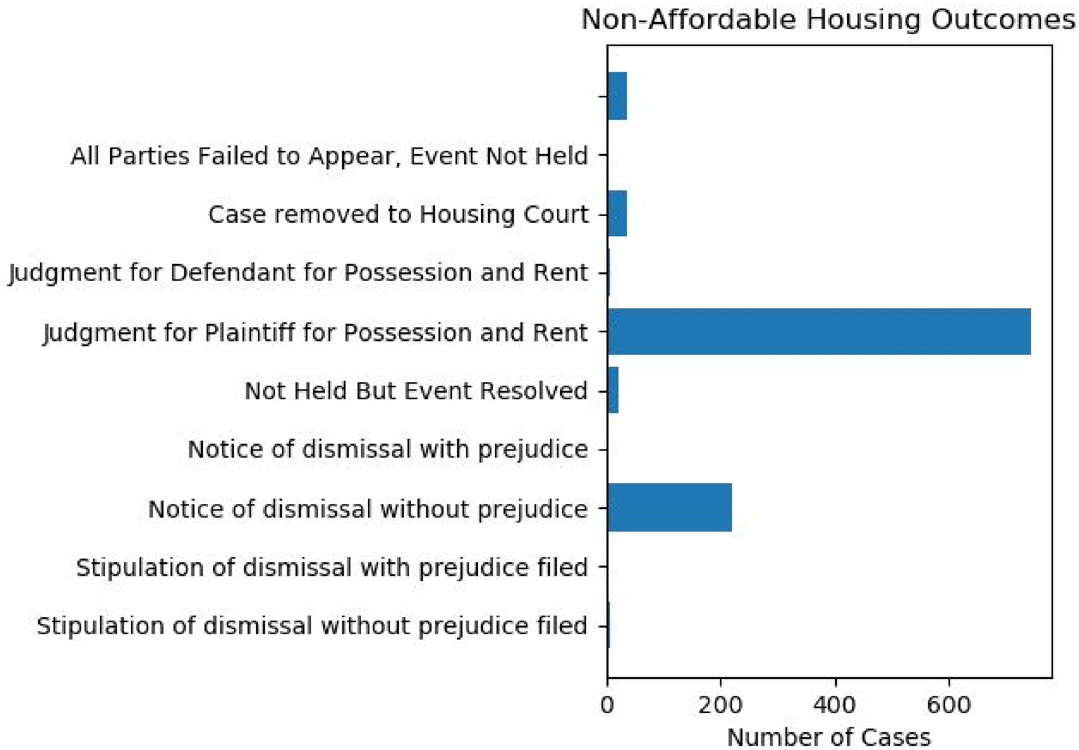
   Another way of analyzing the trends in the characteristics of each defendant case is to cluster data. In order to cluster data, the features were categorized by parsing and categorizing data. Strings, such as neighborhood and judgment type, were categorized by name. Numerical data, such as median income,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

   K-means is 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.

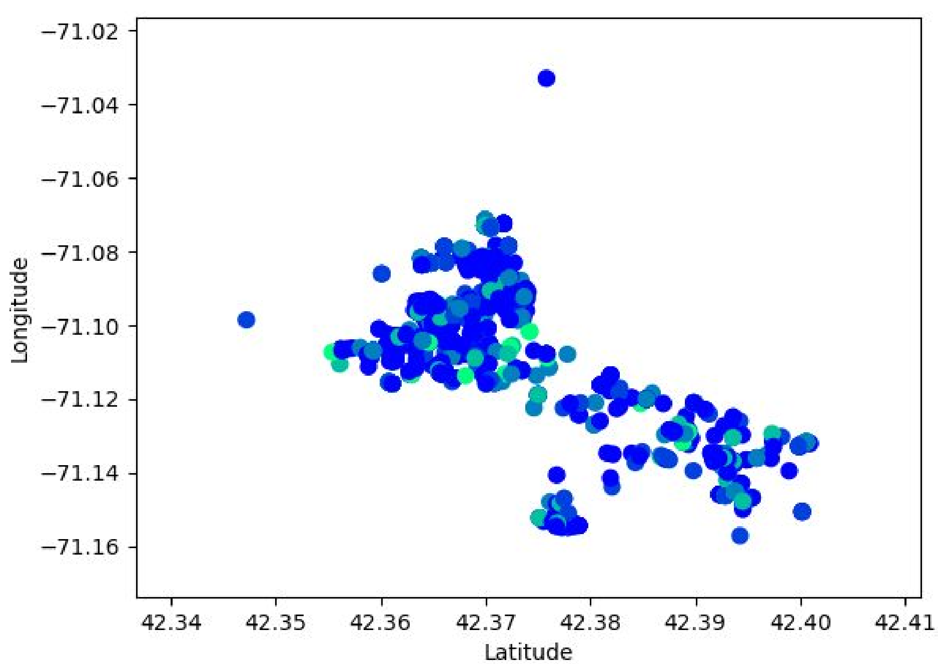
   K-modes was used to examine differences between affordable housing and non-affordable housing. In order to determine the number of low-income or affordable housing units, we used various low-income housing search websites. The addresses in our database were then parsed and searched for the names and 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 eviction for non-affordable housing units compared to affordable housing units.

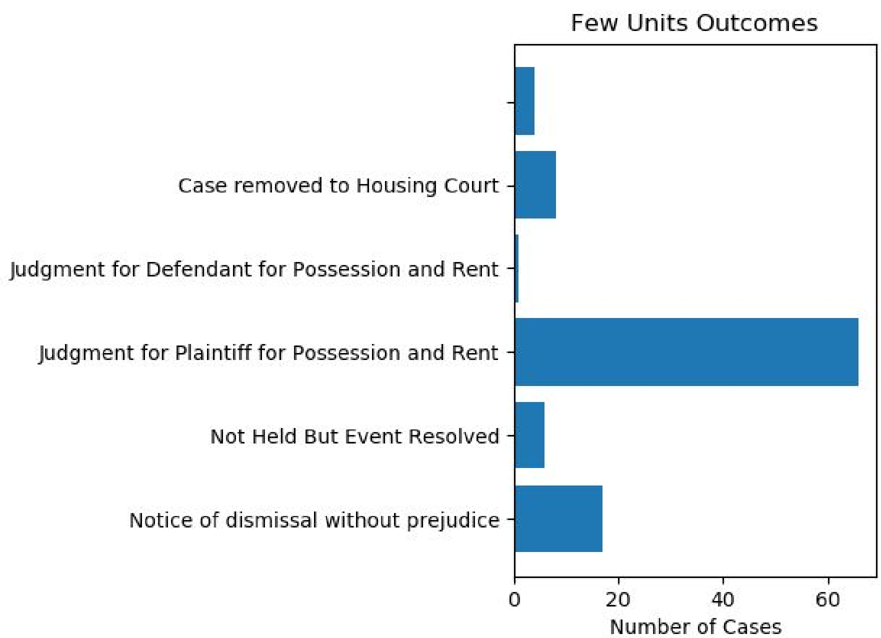


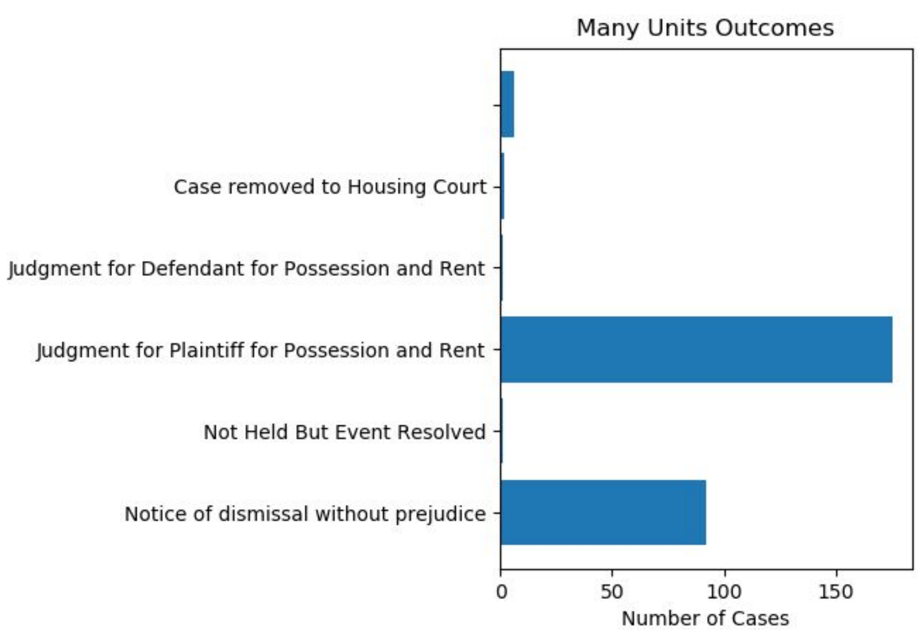


The following figure shows a preliminary K-modes analysis for affordable housing. This may be adjusted and discussed further in the final report.

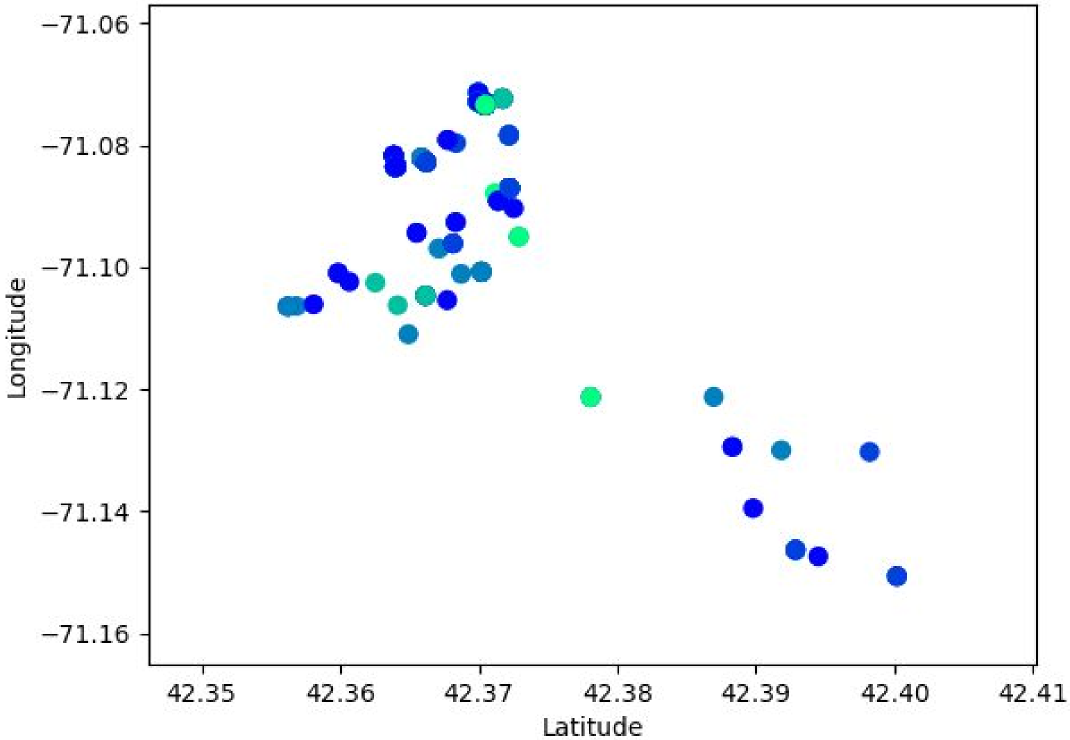


At this point, the data on the size of the apartment units is somewhat limited. We were able to retrieve the number of units at an address for 379 cases (in the final report, we will examine whether we can determine apartment building sizes for any more properties). These cases were then analyzed using 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 than buildings with a small number of units (< 100) compared to the other outcomes.

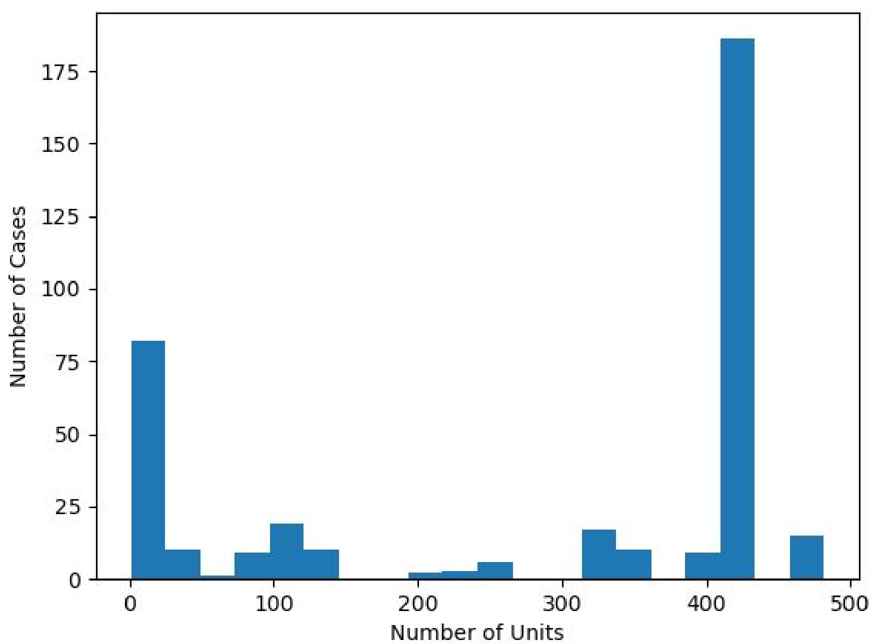




The following figure shows a preliminary K-modes analysis for the number of units. This may be adjusted and discussed further in the final report.



A histogram of the number of units is shown below.

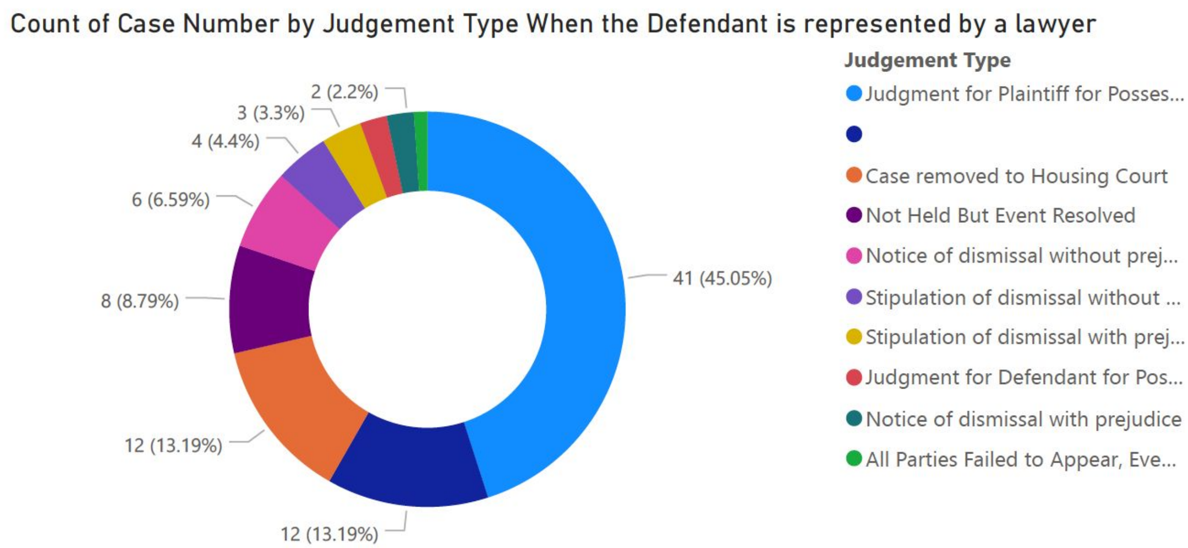


3. How does tenant representation by attorney affect outcomes?

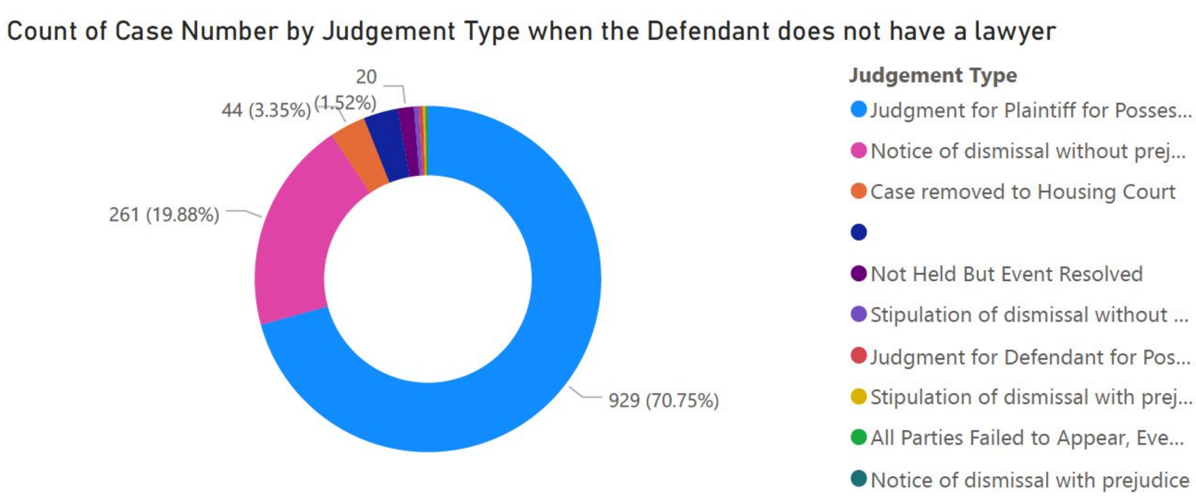
Defendant represented by attorney vs. no representation

Even though there are only 91 cases out of 1404 (less than 6.5%) where the defendant was represented by an attorney, we can see that the outcome for the defendant is better when he was represented by an attorney compared to when he is not represented by an attorney.

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

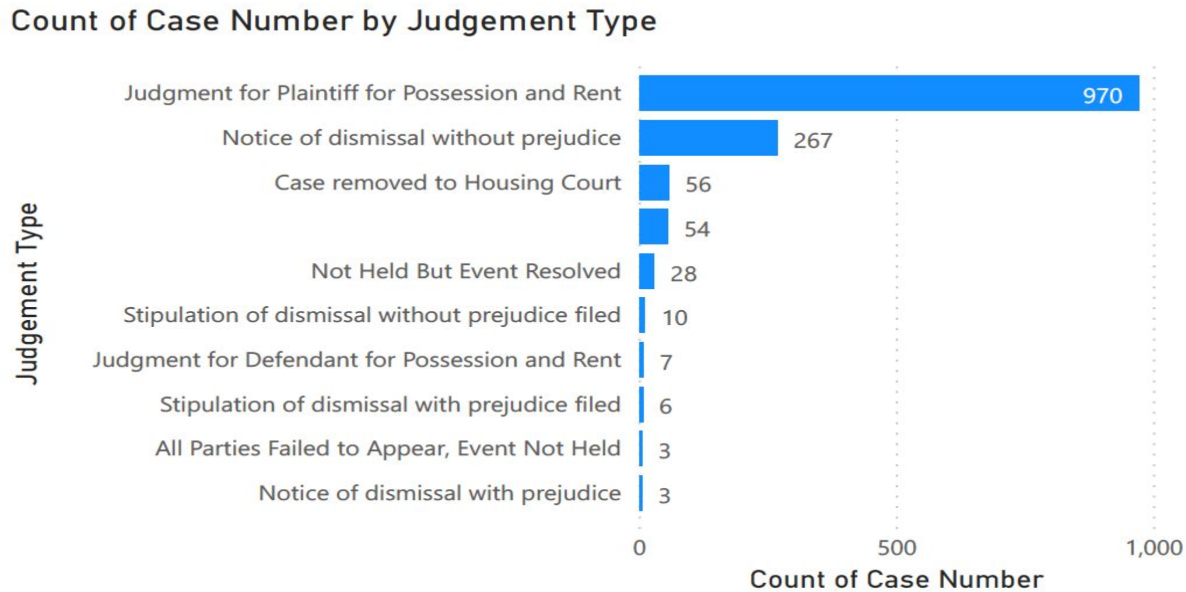


When the defendant was not represented by an attorney, the percentage of cases that resulted in the Judgement for Plaintiff increased to 70% of cases.



Cases went through trials by a judge vs no trials

The super majority of cases (970 out of 1404) resulted in judgement for the plaintiff and 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 the judgement for the defendant increase to 1/7 from overall cases 7/1404.

Figure below is the number of judgement types after trials by a judge.

Additional Data:

Number of units and residential building permits:
https://data.cambridgema.gov/Housing/Housing-Starts-Map/2rxh-46kx

Address/coordinate dataset:
https://data.cambridgema.gov/Geographic-Information-GIS-/Master-Addresses-List/vup6-kpwv

Income by neighborhood:
https://data.cambridgema.gov/Neighborhood-Census-Data/American-Community-Survey-2013-17-Estimates-by-Nei/maef-iias

Affordable Housing Links:
https://www.publichousing.com/city/ma-cambridge
https://www.lowincomehousing.us/MA/cambridge.html
https://affordablehousingonline.com/housing-search/Massachusetts/Cambridge?page=6#apartments