CS 506
Building Violations

Final Report

Team C

Youxuan Ma, Class of 2024, markma@bu.edu
Heyang Yu, Class of 2025, jhyyu@bu.edu
Jian Xie, Class of 2025, jianx@bu.edu
Guanxi Li, Class of 2025, guanxili@bu.edu
Yuzhe Jiang, Class of 2025, jiangyz@bu.edu (Team Rep)

April 2024

Contents

1 Introduction

1.1 Problem Restatement and Analysis

There are various building violations in the Boston area, which seriously affect urban construction and safety, and cause trouble to residents. These violations are scattered across neighborhoods across the city, with varying densities, types, and severity.

The project aims to analyze various types of violations and requests in Boston to identify patterns, systemic issues, and areas needing improvement. The overall goal is to enhance city governance, public safety, and the quality of life for residents by addressing these violations and requests efficiently.

This project is crucial as it directly impacts the living conditions and safety of Boston's residents. By understanding the nature and distribution of violations and service requests, the city can prioritize resources, enforce codes more effectively, and implement preventive measures to reduce future incidents.

1.2 Abstract of Our Works

In order to analyze this project more clearly and answer key questions more accurately, our research is mainly divided into two aspects: basic analysis and extended analysis.

In the basic analysis part, we answered four key questions respectively. By cleaning the given datasets, we removed the noise and retained the data that is convenient for analysis. We use technologies such as Descriptive Analytics, Geospatial Analysis, Trend Analysis, Clustering Analysis, etc., and use visualization methods such as Heatmaps, Bar Graphs and Pie Charts, Python Libraries, etc.

In the extended analysis part, we adopted the earlier modification suggestions and based on the basic analysis further explore the complex relationship between various demographic, environmental, financial factors and their impact on building violations across Boston’s neighborhoods. Understanding these dynamics helps in pinpointing targeted interventions for reducing violations and enhancing urban compliance and safety.

2 Base Analysis

2.1 Data Cleaning and Preprocessing

We found that there are a large number of missing rows and missing columns in the provided data set. The existence of these noise information affects the accuracy of data processing.

1. Remove columns with substantial missing data to reduce interference from irrelevant data.
2. Delete rows with missing information to ensure all data are valid.
3. Sort the dataset to make it more structured and organized.

2.2 Are there certain landlords/ management companies that have repeated violations?

2.2.1 Landlords / Management with Most Violations

By merging the ST_NUM and ST_NAME columns from the Property Assessment dataset, along with the violation_stno, violation_street, and violation_suffix from the Building and Property Violations dataset, we have created full_address columns in both tables. Subsequently, by matching these datasets on full_address and tallying the OWNER column in the resulting matches, we are able to determine the number of violations attributed to each landlord/management entity. This analysis is instrumental in identifying and safeguarding against landlords with a history of violations, thereby protecting potential tenants from undesirable rental situations.

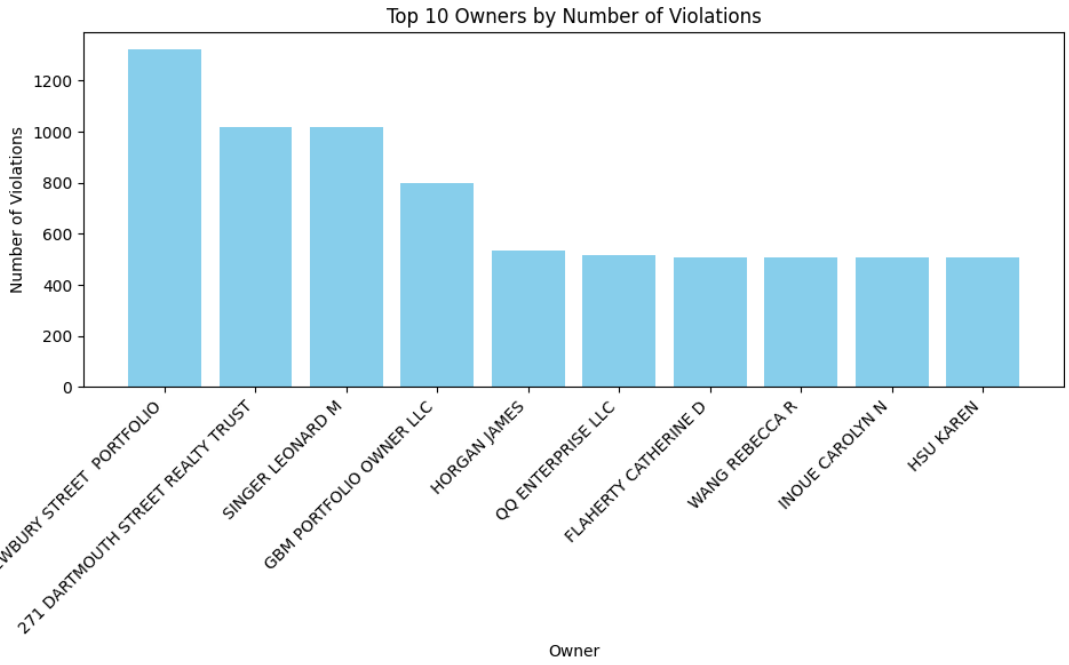

Figure 1: Owners by Number of Violations

As we can see, there are many landlords / management companies that have repeated violations. For instance, Ap Newbury Street Portfolio, 271 Dartmouth Street Reality Trust and Singer Leonard M have significant repeated violations.

After our analysis, we believe that the number of buildings occupied by these landlords/ management companies will also significantly affect this value. I hope it can be handled reasonably in future work.

2.3 Are there common features of certain buildings?

First off, we noticed that the dataset didn't have latitude and longitude coordinates, only detailed street addresses. So, we used geocoding tech to get the latitude and longitude for all the addresses. This will help us visualize some features on a map later on.

To better analyze this issue, we approach it from the perspective of someone who needs to select a house. The first aspect we focus on is the type of housing. Therefore, we conducted visual analysis for each category of housing type, obtaining the actual distribution and quantity for each type. This enables us to clearly understand where suitable housing types are located if I need to work or study in a certain area.

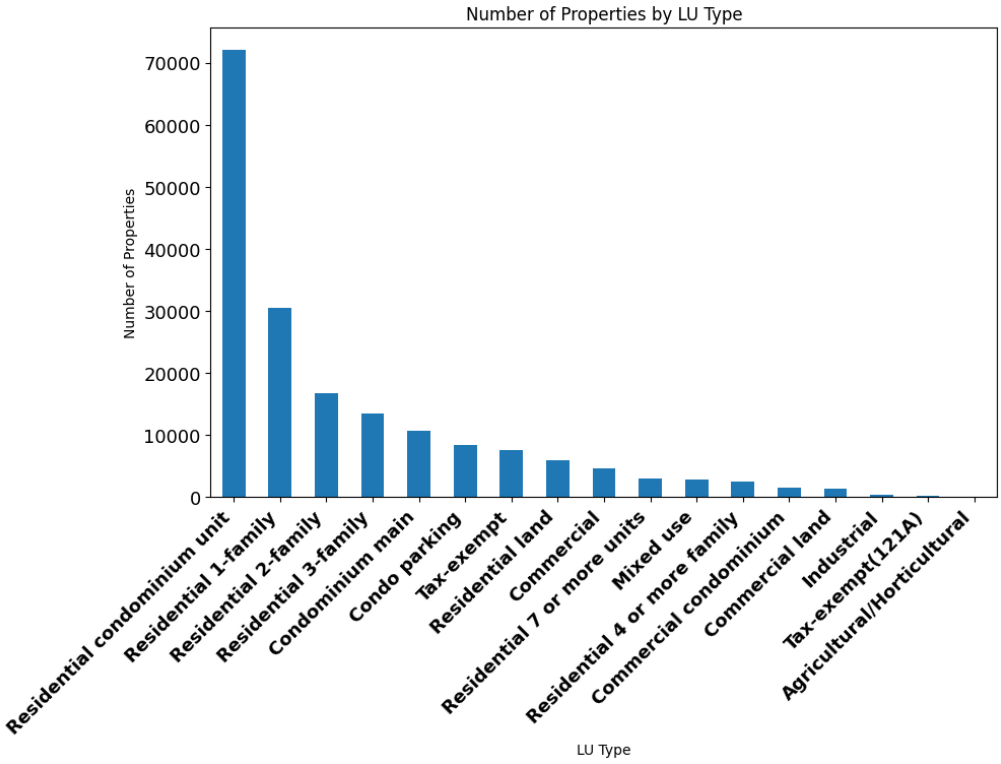

Figure 2: Number of Properties by LU Type

We analyzed the PROP_VIEW column and visually analyzed properties with different ratings, including their distribution on the map and quantity comparisons. Among them, the majority have an average rating, while only a few of them have a poor rating. We also performed similar clustering on PROP_VIEW to display the distribution of different ratings.

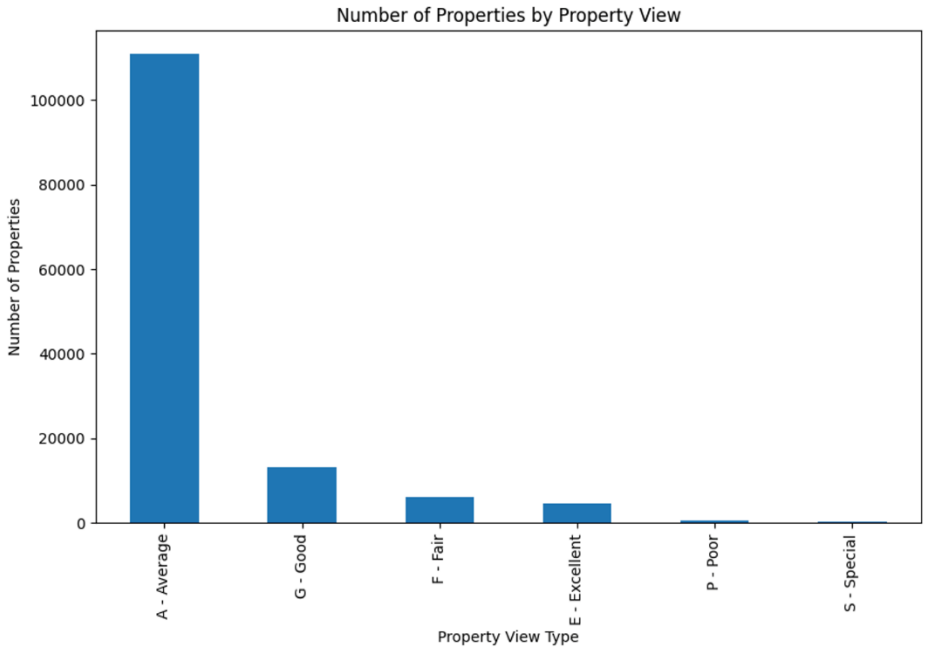

Figure 3: Number of Properties by Property View

Next, we analyzed the indoor and outdoor conditions. Considering that different people have different priorities, we designed three weighting methods to rate the indoor and outdoor conditions of the property. The three ratings take into account those who are more concerned about indoor environments, those who focus more on outdoor environments, and those who prioritize overall conditions.

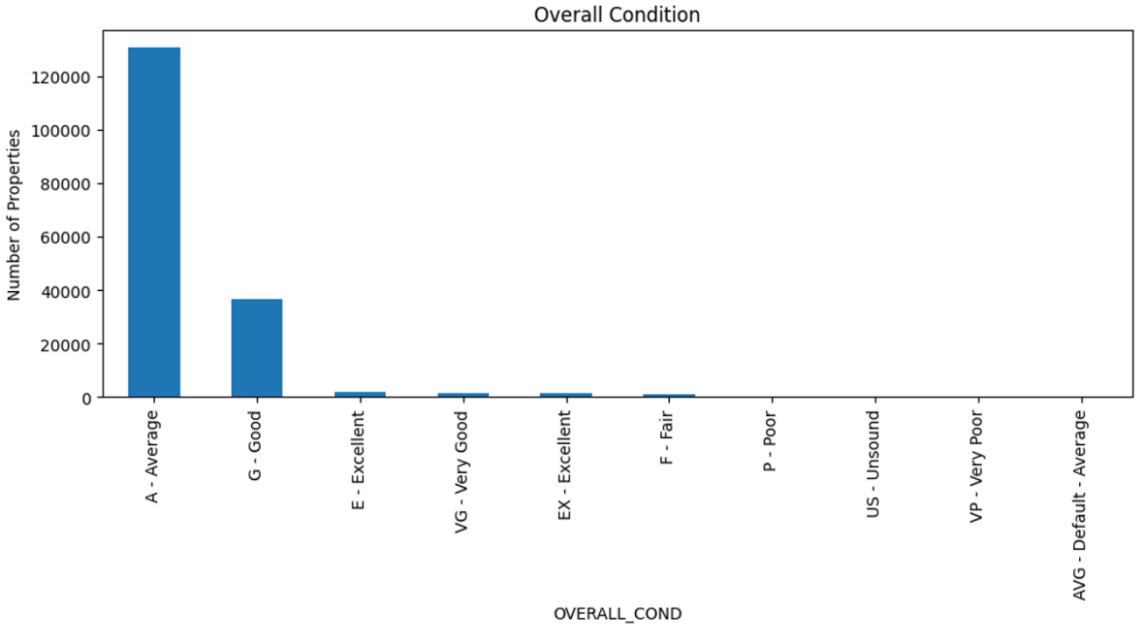

Figure 4: Overall Condition

Finally, considering that many people are concerned about the age of the house, we explored the construction year.

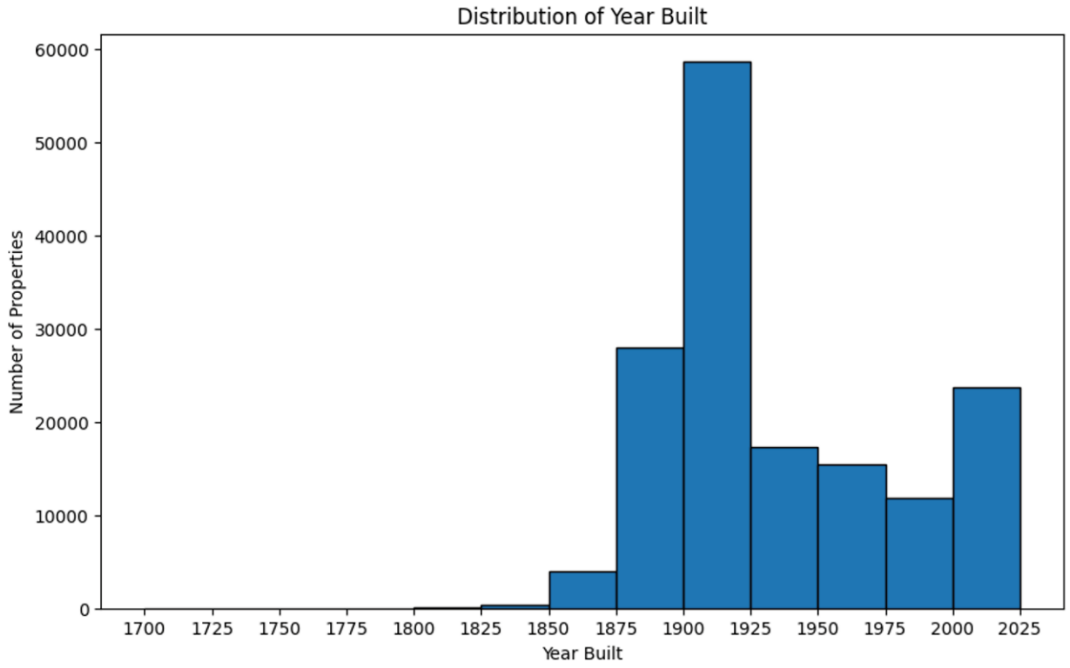

Figure 5: Distribution of Year Built

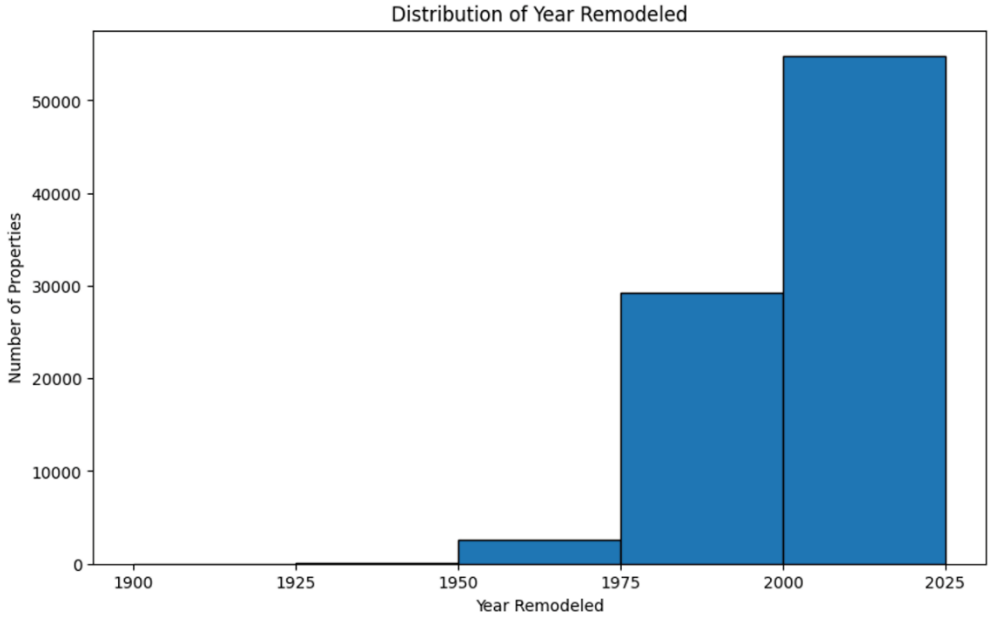

Figure 6: Distribution of Year Remodeled

2.4 What neighborhoods / communities are affected most?

2.4.1 Top 20 Streets with the Highest Number of Violations

By counting the occurrences of `violation_street` entries, this chart ranks the top 10 streets with the highest number of reported violations. This visualization helps to pinpoint hotspots of non-compliance within the area, shedding light on where resources might be best allocated to address these issues.

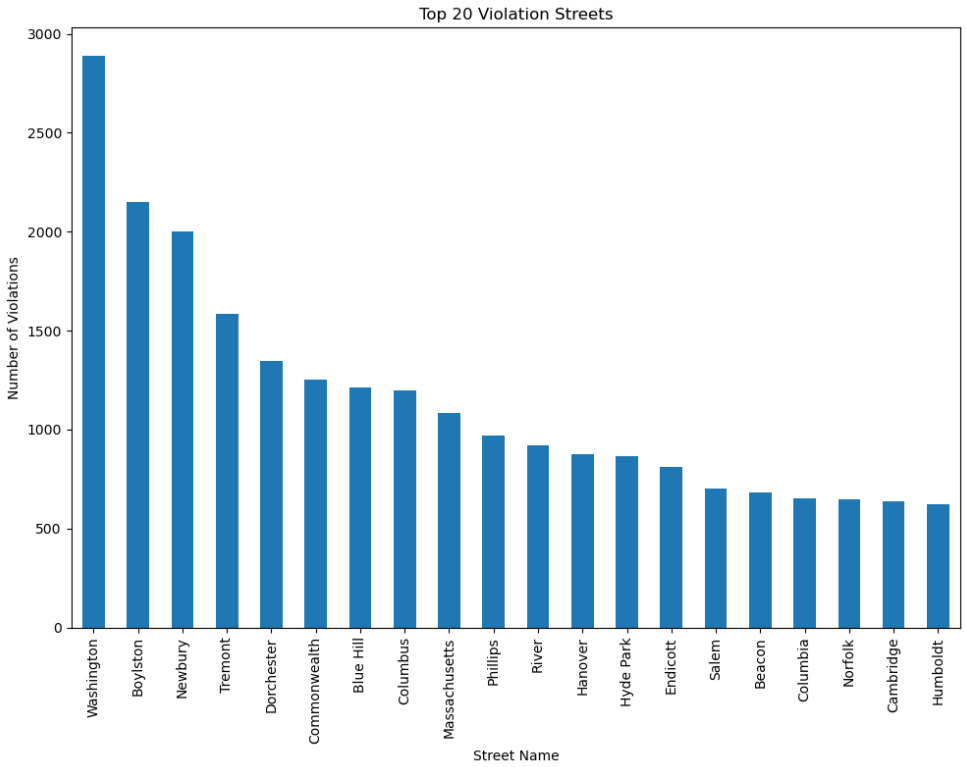

Figure 7: Top 20 Violation Streets

2.4.2 Streets with Repeated Violations

This visualization identifies streets with recurring violations by aggregating records based on the `sam_id`, a unique identifier for properties. It filters for `sam_id` occurrences greater than two and then counts how frequently each corresponding street is mentioned. This graph highlights streets prone to repeated violations, suggesting areas that might benefit from targeted enforcement or preventative measures.

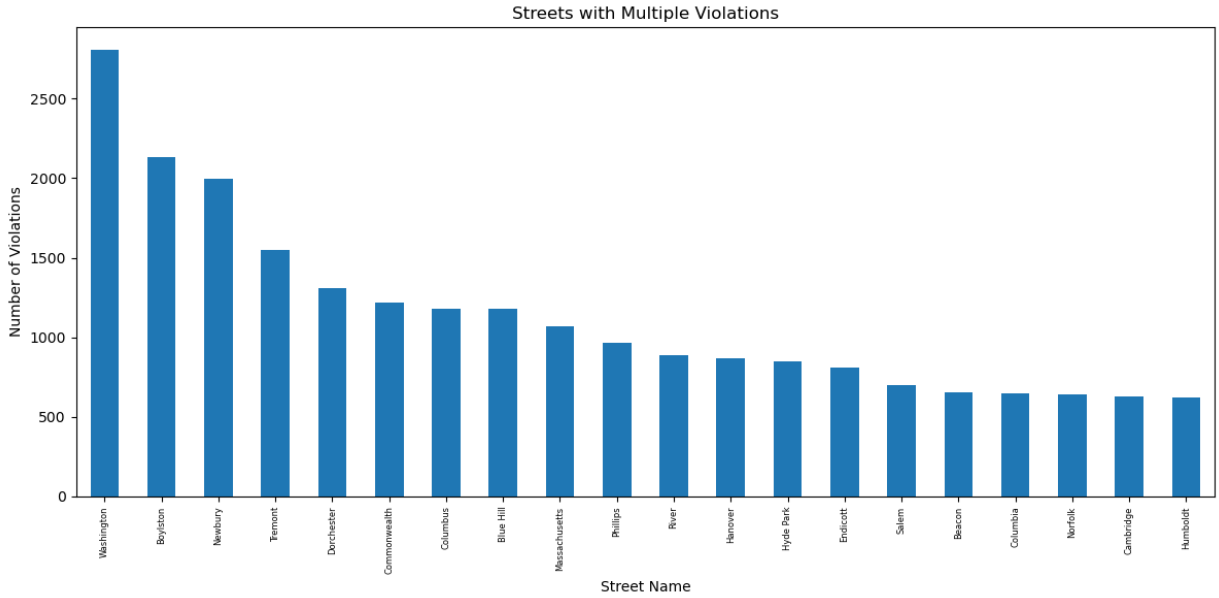

Figure 8: Streets with Repeated Violations

2.4.3 Data analysis results applied to the map

Based on the latitude and longitude of each record, this heatmap visualizes the concentration of violations in specific geographic areas. Areas with a higher density of violations appear more prominently, allowing for a quick visual assessment of which neighborhoods might be experiencing higher levels of non-compliance or other issues.

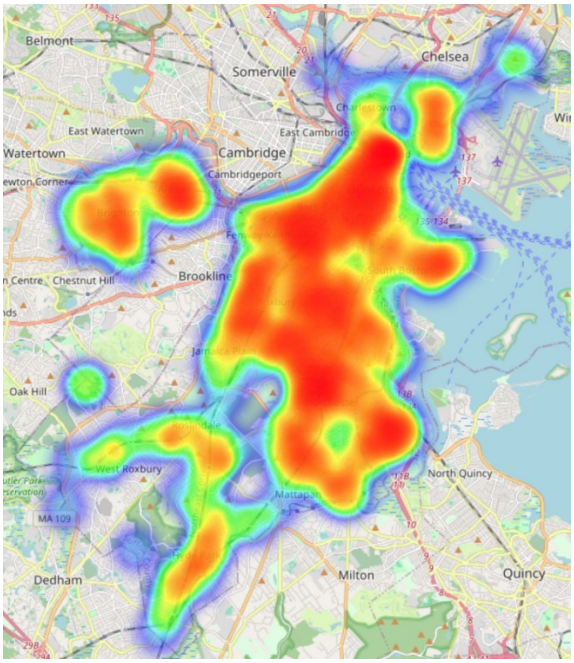

Figure 9: Heatmap on real map

This map marks the exact locations of recorded violations using the geographic coordinates provided in the dataset. It offers a detailed view of where violations are occurring within the city, enabling a granular analysis of problem areas. This visualization can be particularly useful for local authorities or urban planners looking to address specific issues or improve citywide compliance strategies.

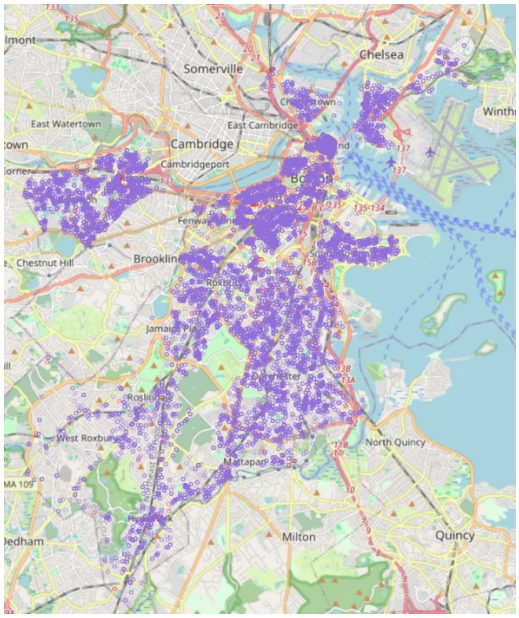

Figure 10: Boston Violations Map

2.4.4 Most Affected Neighborhoods

These two charts count the ‘neighborhood’ column in the data from 2020 onwards for the ‘Building and Property Violations’ and ‘Public Work Violations’ datasets, respectively. The goal is to identify which neighborhoods have been most severely affected by violations in recent years.

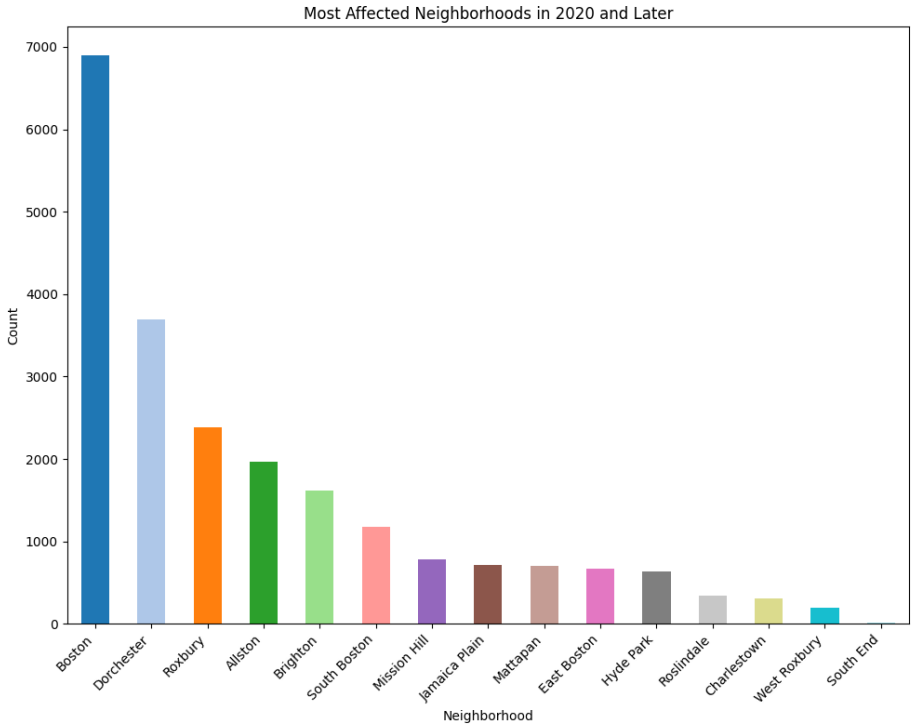

Figure 11: Most Affected Neighborhoods based on Building and Property Violations

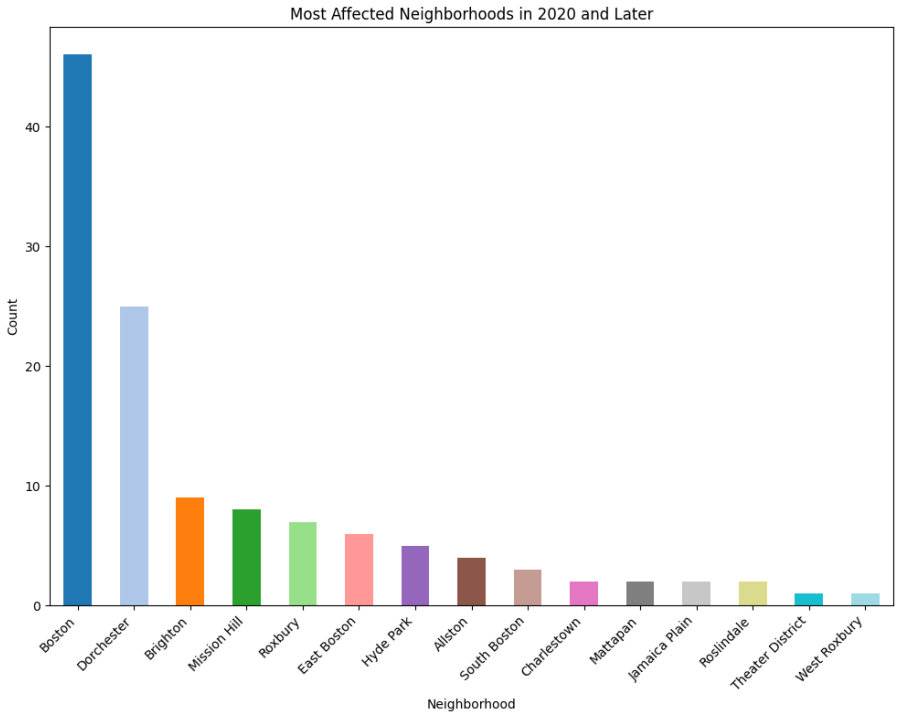

Figure 12: Most Affected Neighborhoods based on Public Work Violations

As we can see above, these neighborhoods / communities are affected seriously. In addition to the dendrogram, we also plotted heat maps to visualize the distribution of building violations in Boston.

2.5 What kinds of building complaints are people making around the city?

Utilizing the `description` field, this heat map displays the frequency of each type of violation across different streets. It reveals patterns in the prevalence of specific violations per street, providing insights into common compliance issues in certain areas. This information could guide targeted interventions or public awareness campaigns.

We can see the density of different types of building violations on different streets and be able to work out which violations are most prevalent in a particular area. Additionally, this report shows an abbreviated version because the full version of the heat map is too large.

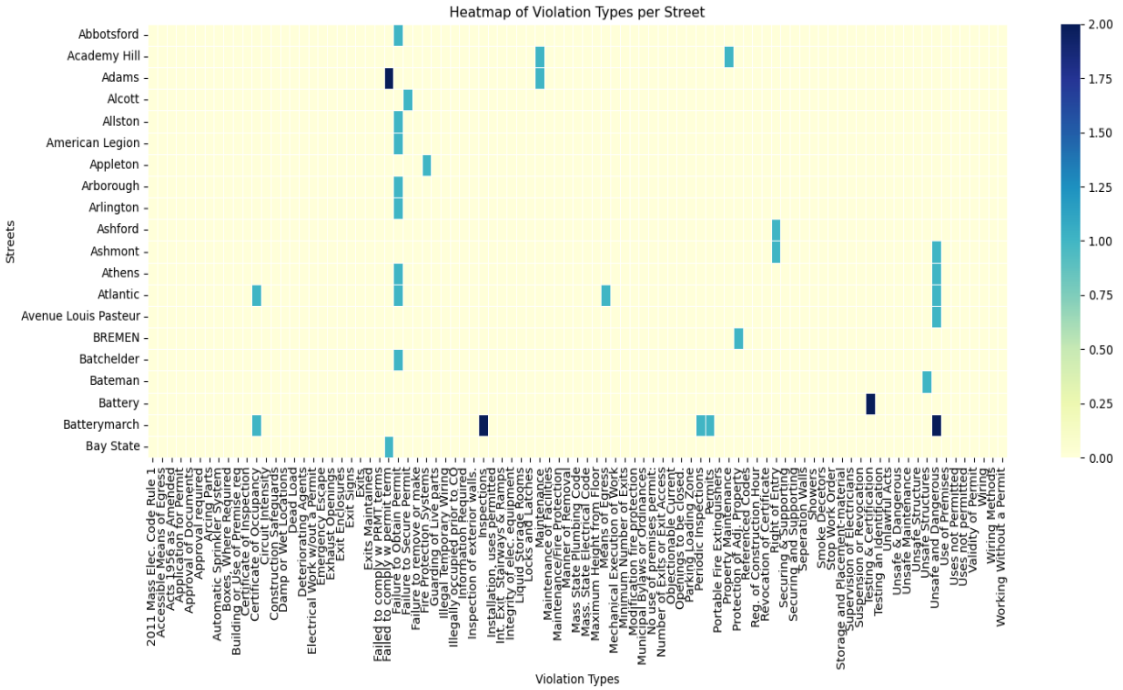

Figure 13: Heatmap of Violation per Street

We refined our categorization of issue types and ended up with 40 different types of issues. To really understand what each of these 40 types means, we looked for data points that best represent the center of each cluster. We did this by finding the point closest to its cluster's center. This way, we got a clear description for each type of issue.

However, 40 types of issues were still too many for further analysis, so we used a Large Language Model to categorize these 40 types into 7 groups.

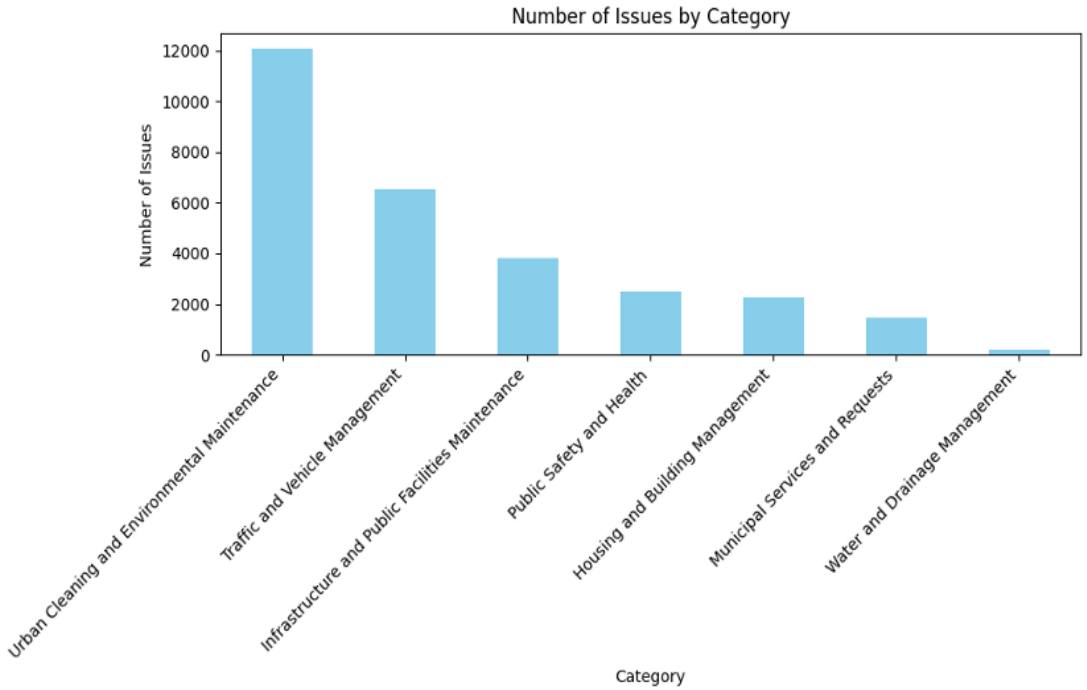

Figure 14: Number of Issues by Category

As we can see, the issues mainly centered around: Urban Cleaning and Environmental Maintenance, Traffic and Vehicle Management, Infrastructure and Public Facilities Maintenance

3 Extension Analysis

We want to explore the complex relationship between various demographic and environmental factors and their impact on building violations across Boston's neighborhoods. Understanding these dynamics helps in pinpointing targeted interventions for reducing violations and enhancing urban compliance and safety.

3.1 Population Size vs. Violations

Intuitively, one would expect neighborhoods with larger populations to have more violations. Hence, we first explored the relationship between population size and the number of violations per neighborhood. Since the most accurate population data is only available up to 2020, we analyzed the relationship between violations recorded in 2020 and the population data of that year.

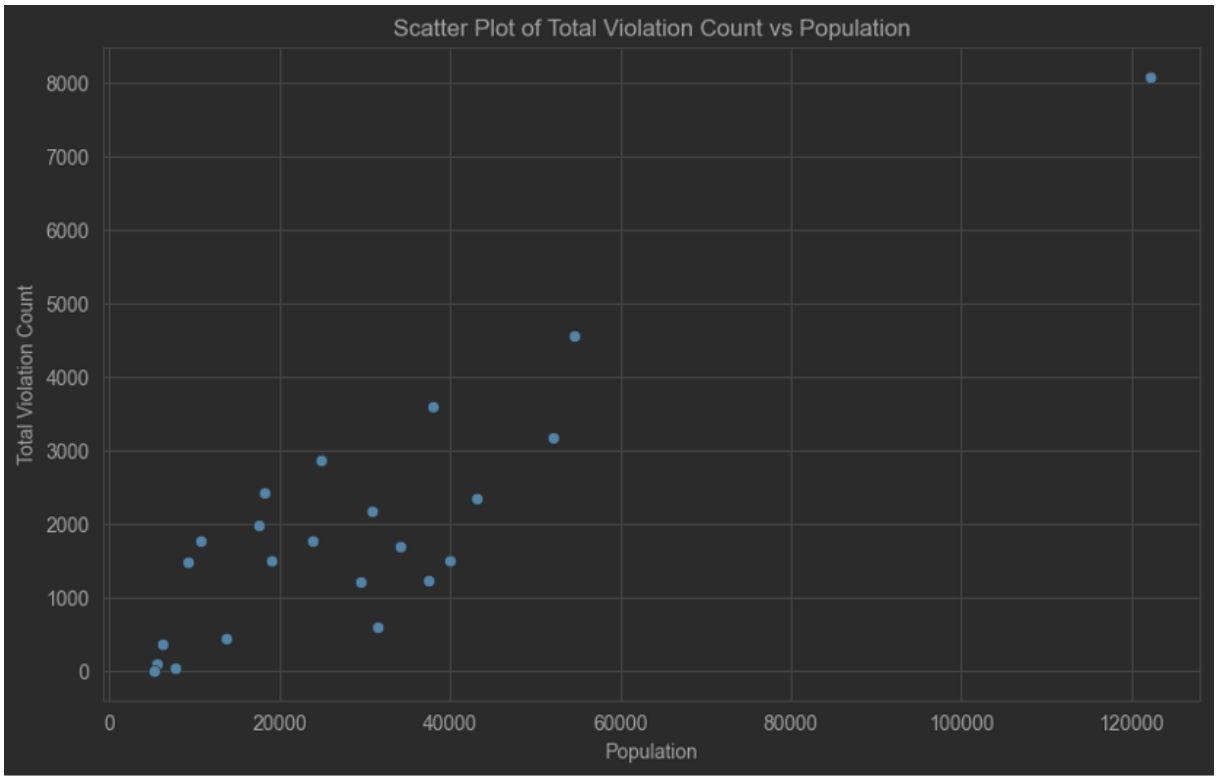

Figure 15: Population vs. Violations

The initial scatter plot confirmed a general positive correlation between population size and the number of violations.

However, using linear regression, and revising the y-axis to Violation Frequency, we found that this relationship is not significant, with no obvious trend and many data points falling outside the 95% confidence interval.

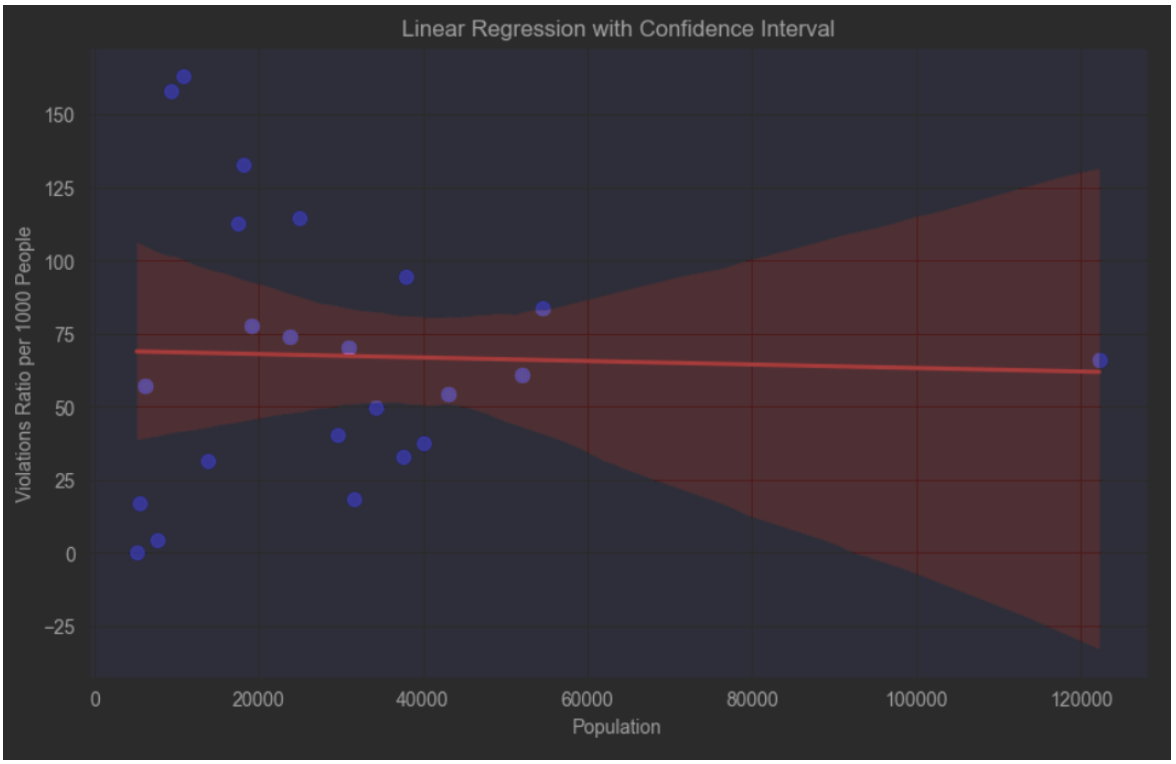

Figure 16: Linear Regression of Population vs. Violations

Thus, population size is not a definitive factor in predicting violations.

3.2 Education Level vs. Violation

Since using population size to predict violations is insufficient, we then delved deeper into the relationship between the proportion of residents with at least a bachelor's degree ("well-educated" population) and violations, using heatmaps and linear regression results.

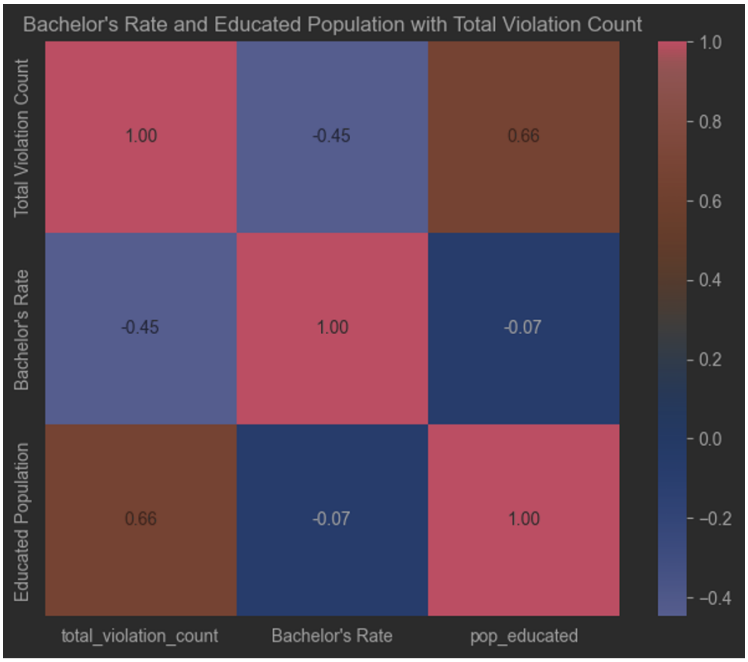

Figure 17: Bachelor's Rate and Educated Population with Total Violation Count

From the graph, there is a negative correlation of -0.45 between the proportion of well-educated residents and violations.

Next, employing linear regression, over 50% of the data points fall within the 95% confidence interval, indicating a certain negative correlation between education level and violations.

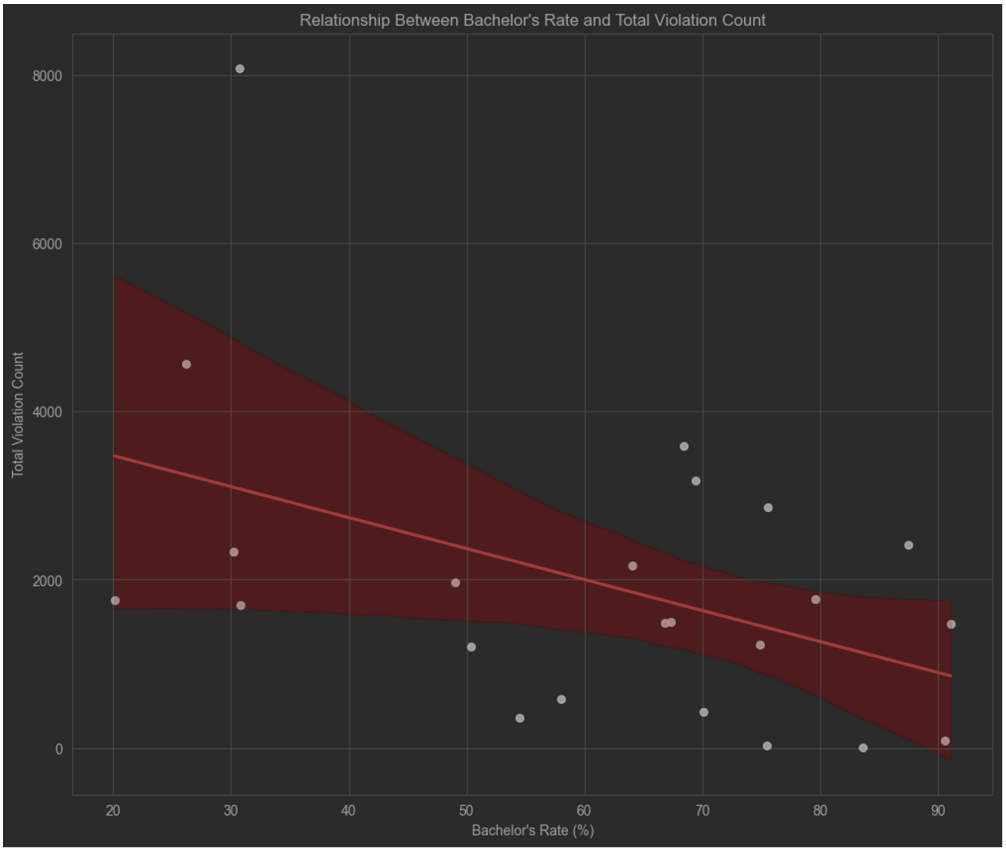

Figure 18: Linear Regression

We also draw two subsequent graphs. They illustrate the bachelor's degree rate by neighborhood, showing Beacon Hill with the highest rate exceeding 85%, while Mattapan has the lowest, around 20%. Another graph displays the number of violations per capita by neighborhood.

Notably, despite Beacon Hill having the highest education rate, it also has the second highest number of violations per capita, suggesting that merely increasing a neighborhood's educational level might not significantly reduce violations.

Conversely, Dorchester, with the highest population and number of violations, ranks below fourth from the bottom in terms of education level.

These indicate that violations are also closely related to many other factors.

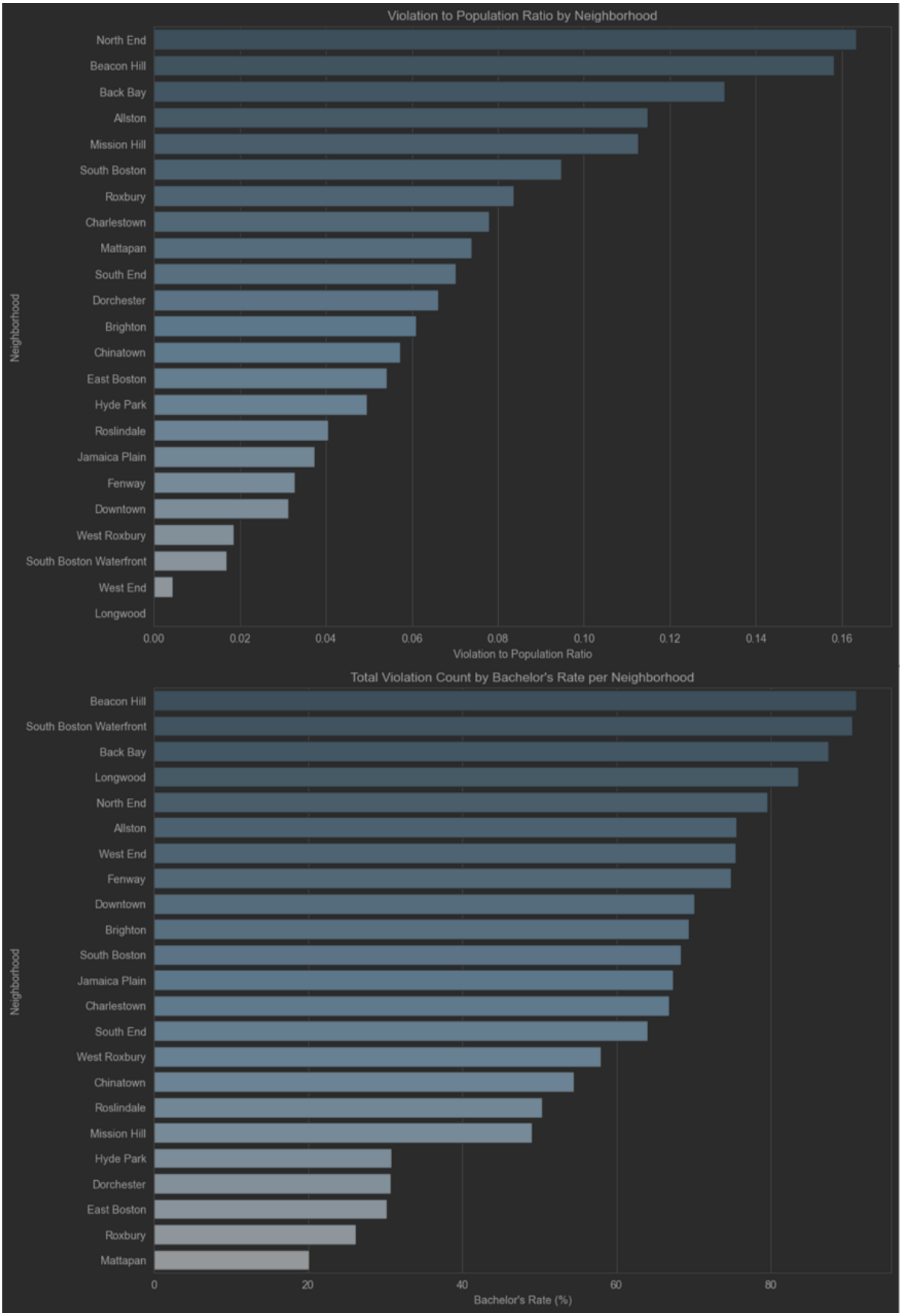

Figure 19: Subsequent Graphs about Education Rate and Violation Rate

3.3 Time-based Analysis of Violations

Since the types of violations in the two datasets are not exactly the same, with each emphasizes different aspects, our subsequent analyses were conducted separately for each dataset. We selected the top 10 most common types of violations from each dataset for analysis.

The BUILDING_AND_PROPERTY_VIOLATIONS dataset focuses on property owners' responsibilities, such as maintaining property cleanliness and compliance with city ordinances, as well as adhering to regulations concerning property use.

The PUBLIC_WORKS_VIOLATIONS dataset is more aligned with public safety and structural integrity, as well as ensuring that building owners follow construction standards and government-imposed procedures.

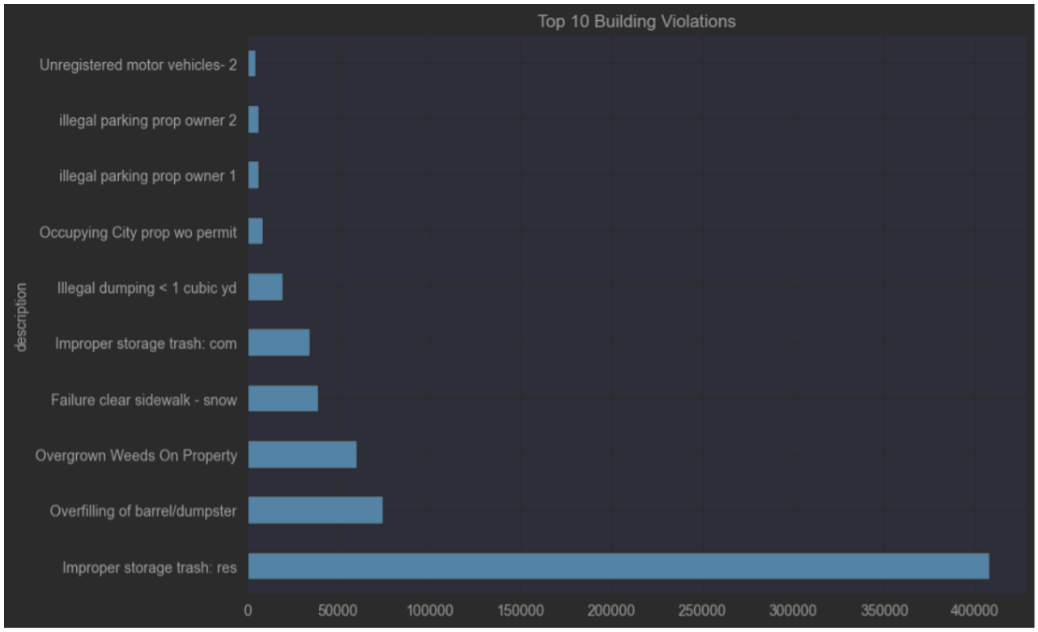

Figure 20: Top 10 Building Violations

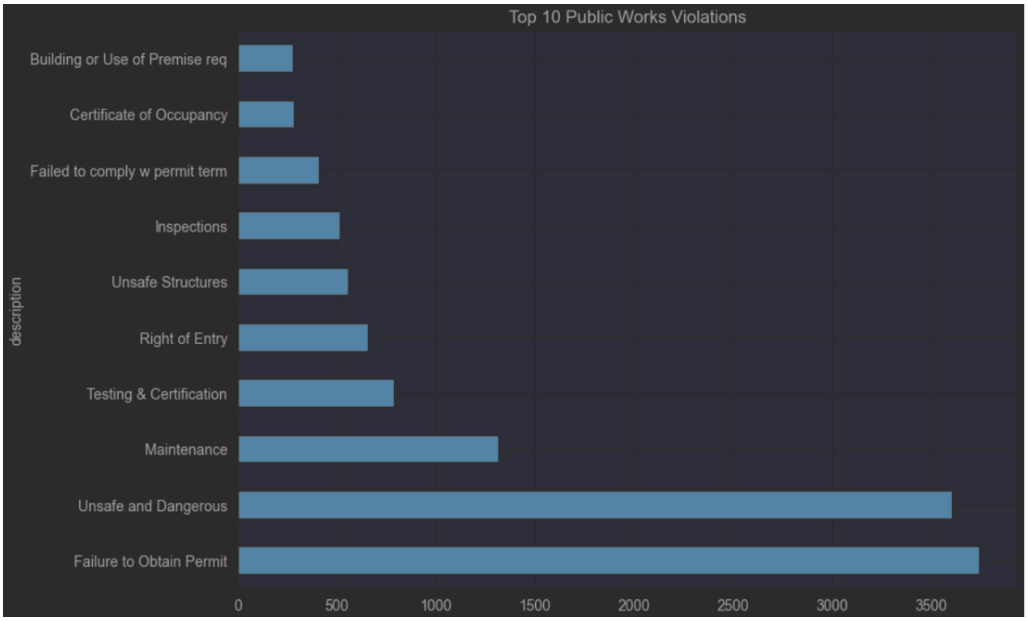

Figure 21: Top 10 Public Works Violations

For the PUBLIC_WORKS_VIOLATIONS dataset, we recorded the number of violations per neighborhood annually and observed trends from around 2010 to 2024. Each neighborhood showed a trend of initially rising and then declining, resembling a hill, but the peak years varied. For example, Brighton peaked in 2014, after which the situation improved.

However, recently in some places, violations have shown an upward trend, like Fenway. This could suggest an increased public safety awareness.

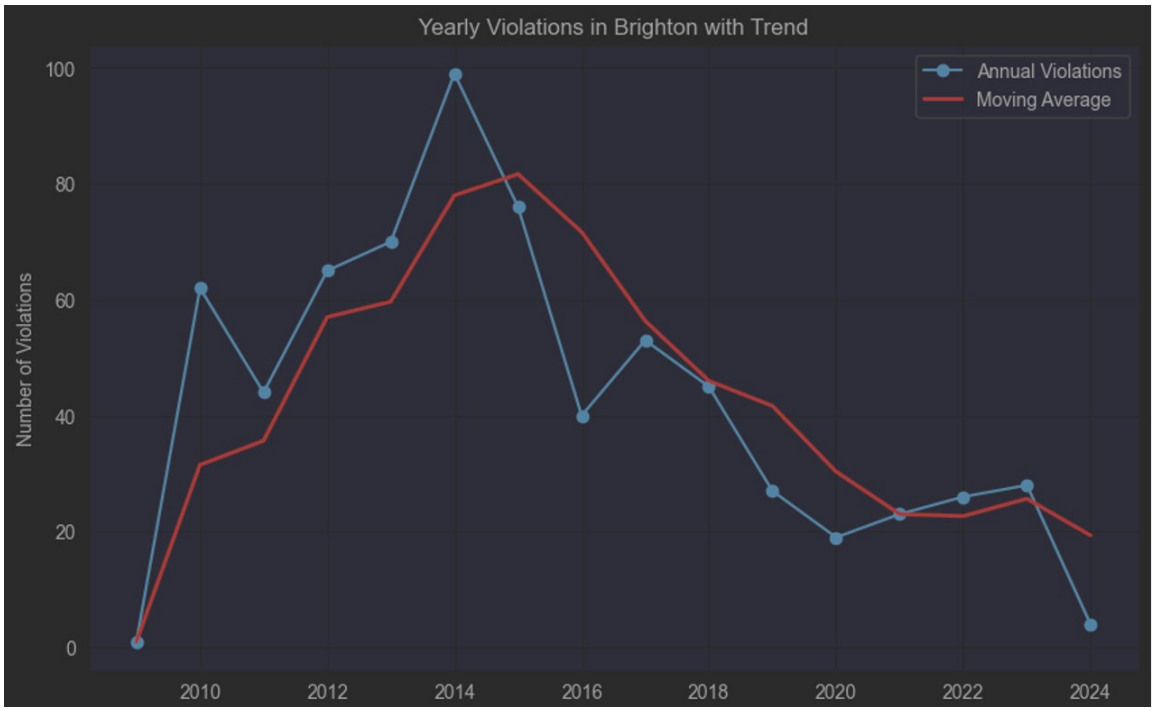

Figure 22: Yearly Violations in Brighton with Trend

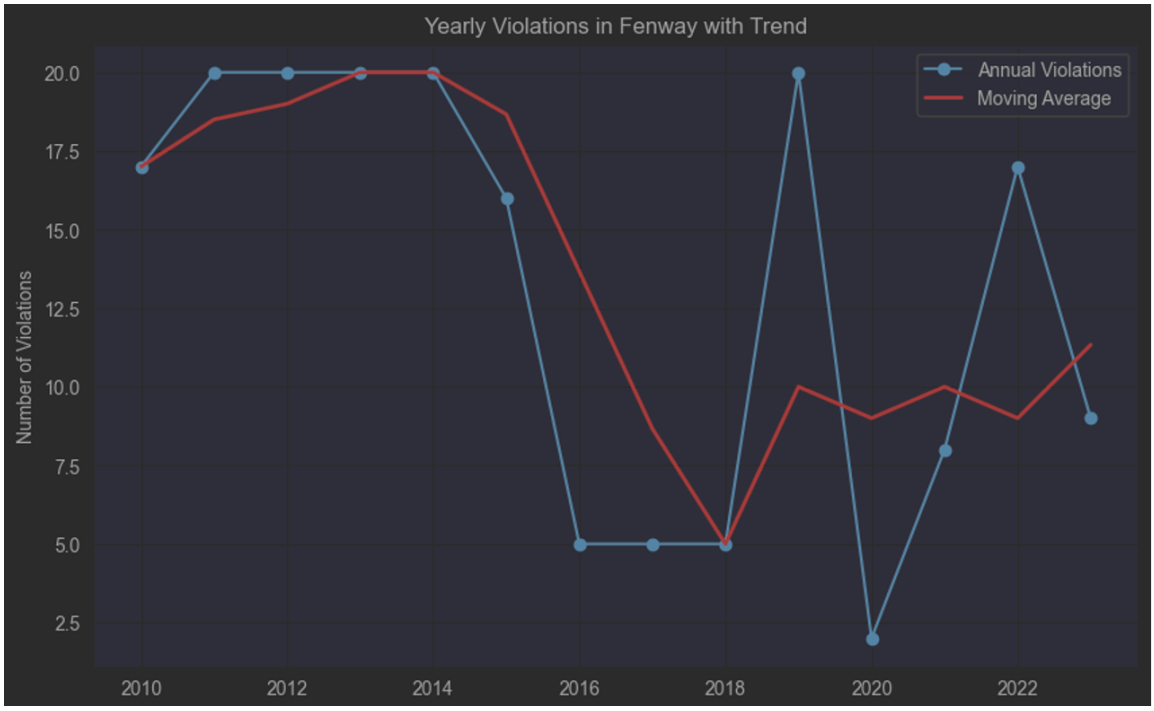

Figure 23: Yearly Violations in Fenway with Trend

Next, we also explored the relationship between seasons and violations, representing each neighborhood's quarterly data annually with different colors on bar graphs. Each neighborhood

displays unique characteristics, and there is no consistent trend applicable to all neighborhoods.

For instance, shown by the graphs, violations in Allston are mainly centered around the third quarter, whereas in Brighton, it's basically the same across the second, third, and forth quarter.

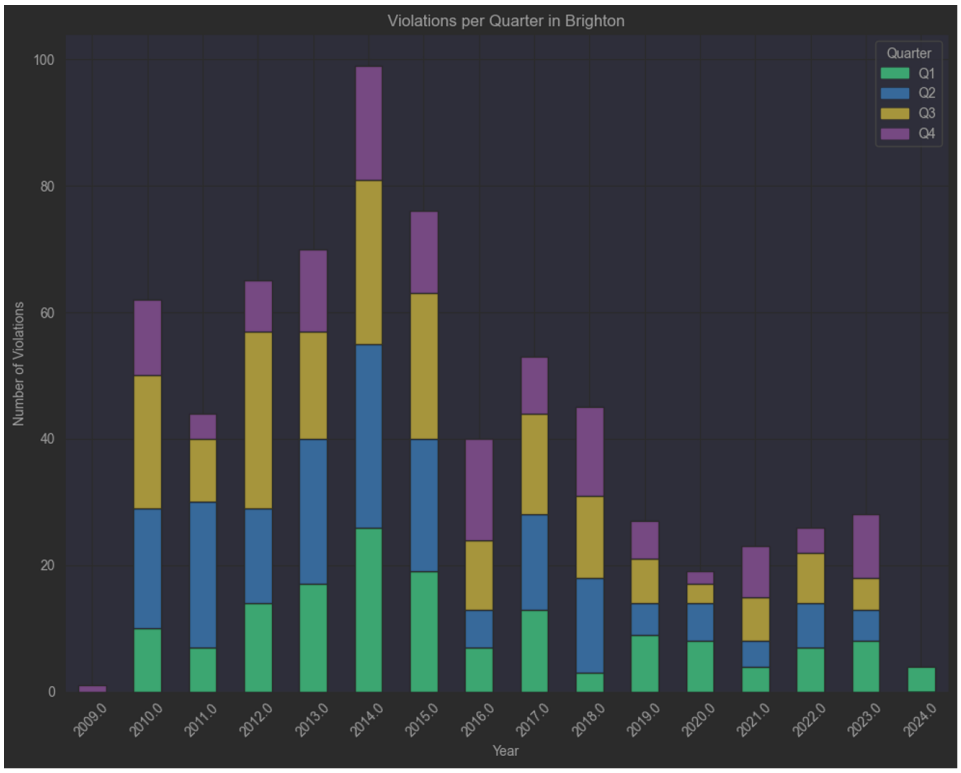

Figure 24: Violations per Quarter in Brighton

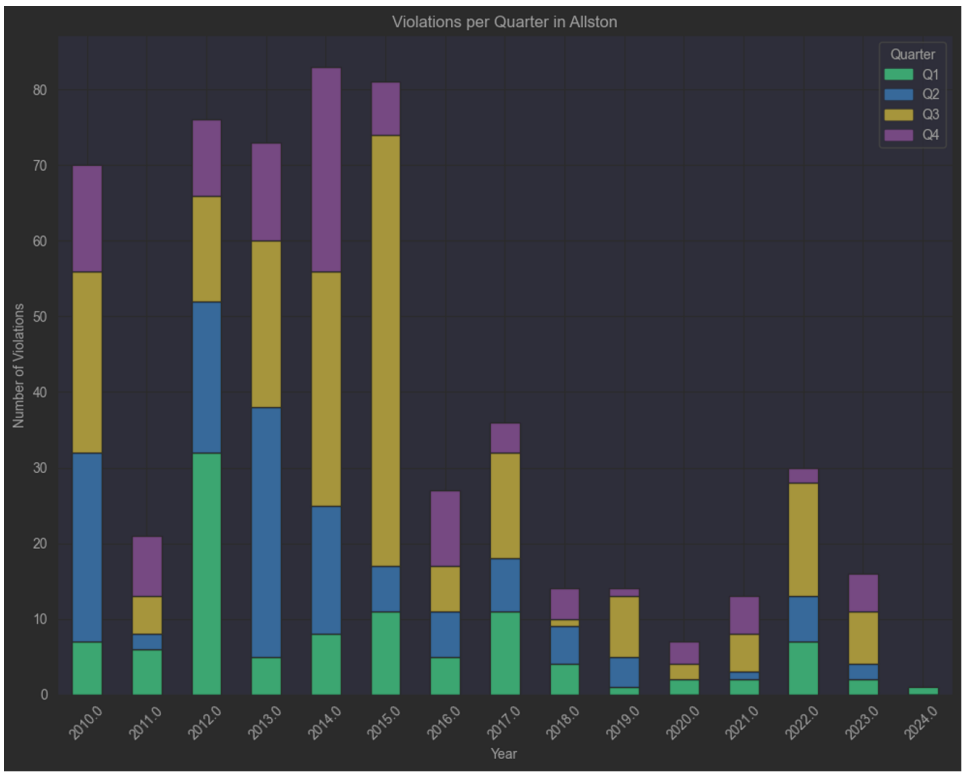

Figure 25: Violations per Quarter in Allston

Lastly, we explored the overall case by quarter. The box plot beside shows that the second and third quarters have slightly higher median violation numbers and quartile ranges than the first and fourth quarters, indicating that these quarters see more violations.

From the trend lines of each quarter annually, overall violations are declining, but each quarter has had years when it led in numbers, showing that no certain quarter is consistently prone to more violations.

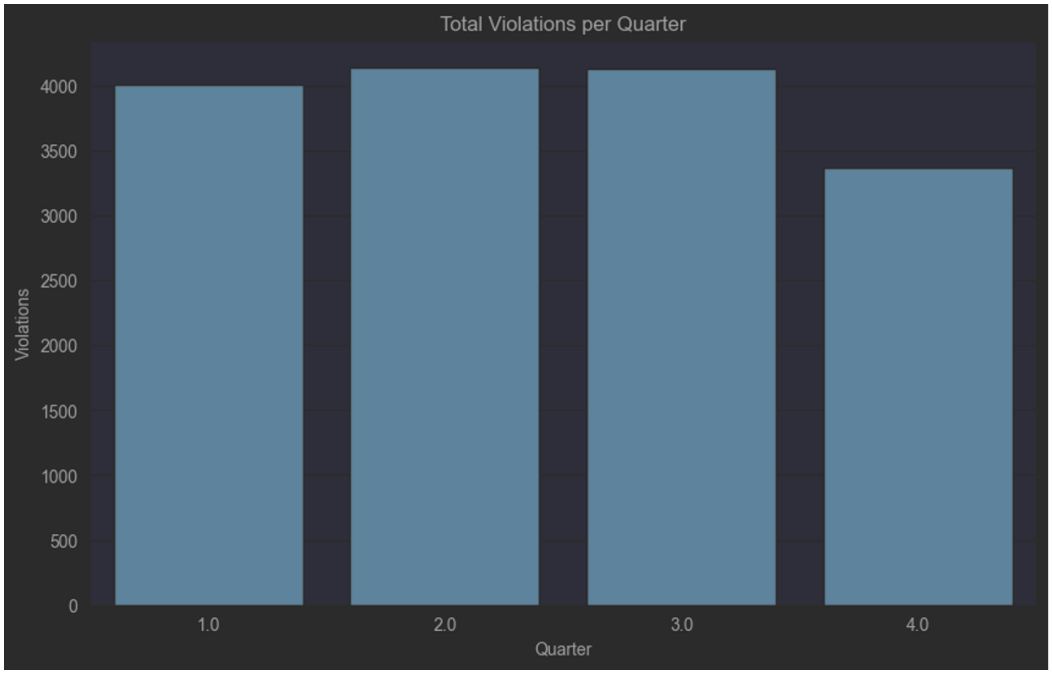

Figure 26: Total Violations per Quarter

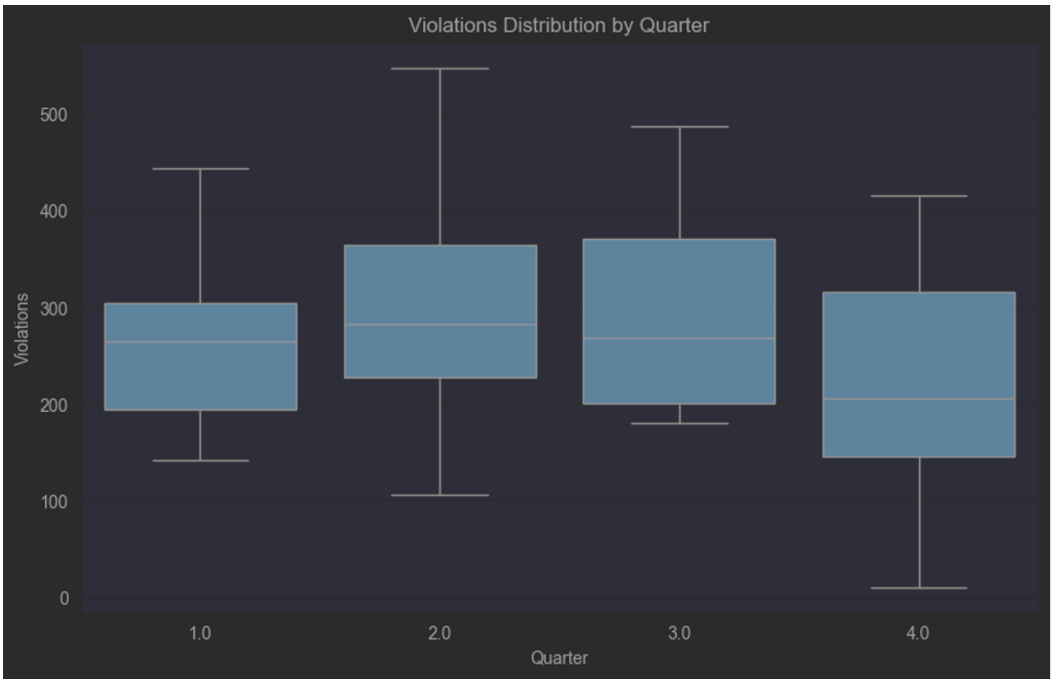

Figure 27: Violations Distribution by Quarter

Figure 28: Trend of Violations Over Years by Quarter

Then, for the BUILDING_AND_PROPERTY_VIOLATIONS dataset, we did similar analysis as before.

While in the last dataset, most neighborhoods generally show a declining trend of violations in recent years, this dataset indicates an increasing trend in the number of violations across most neighborhoods from 2015 to 2020. Specifically, Allston and Fenway have seen a steady increase in violations over recent years.

This trend could likely be associated with changes in regulatory enforcement or other external factors affecting these areas.

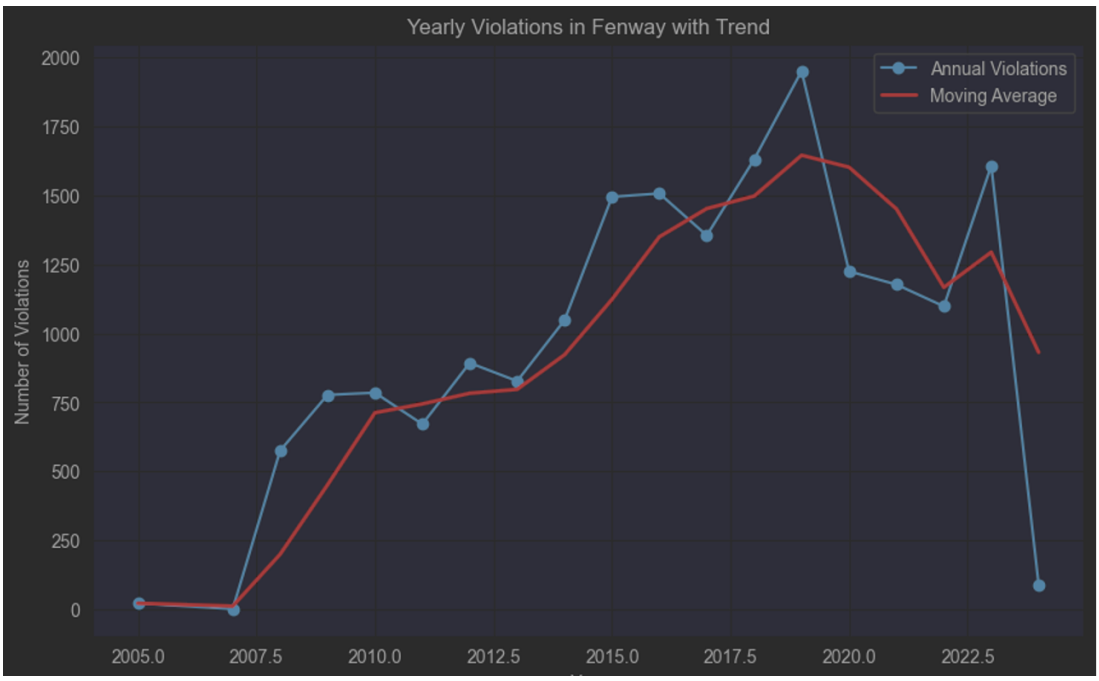

Figure 29: Yearly Violations in Fenway with Trend

Figure 30: Yearly Violations in Allston with Trend

Next, we also explored the relationship between quarters and violations. The box plot and bar graph below show that the number of violations in the third quarter exceeds the other three quarters.

This suggests that issues such as property maintenance, cleanliness, and compliance with city ordinances are more likely to occur in the third quarter.

We think that this may be closely related to the moving season associated with the start of the academic year, especially since Boston is a city with a high concentration of universities.

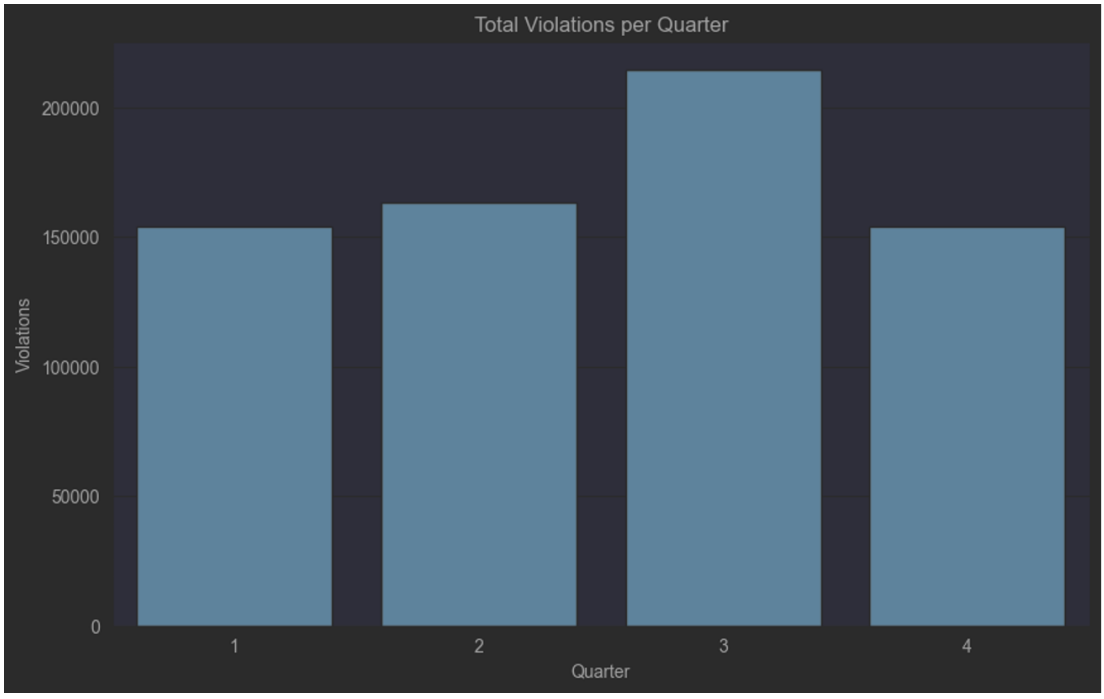

Figure 31: Total Violations per Quarter

Figure 32: Violations Distribution by Quarter

Finally, from the line graph, it is evident that the third quarter, represented by the green line, is gradually becoming the quarter with the most violations. This indicates that the third quarter is becoming the peak season for violations within the year.

Figure 33: Violations Distribution by Quarter

3.4 Economic Influences on Violations

Extending our analysis further, the following segment of our study focuses on deciphering the economic dimensions that may influence the incidence of building violations in Boston. We investigate the interconnections between rental costs, resident income levels, and their collective impact on housing violations.

Our primary hypothesis posits that higher rent indices correlate with better housing quality, hence lower violation frequencies. Conversely, we surmise that lower household incomes could be associated with increased violation rates due to potential constraints on housing maintenance and quality of living conditions.

We sourced our rental data from Zillow's Observed Rent Index (ZORI), which provides a comprehensive view of the typical market rent. For income statistics, we utilized data from incomebyzipcode.com. These economic indicators were cross-referenced with housing violation frequencies derived from Boston's 311 service requests dataset.

Our linear regression model, displayed on the graph, reveals a downward trend line. This negative slope indicates that as ZORI increases, suggesting higher rents and potentially better housing quality, the frequency of violations per 1,000 units tends to decrease.

While the overall trend supports our hypothesis, the presence of outliers signifies the influence of additional factors. Some high-rent districts still experience a notable number of violations, indicating that high rent alone isn't a blanket deterrent to violations.

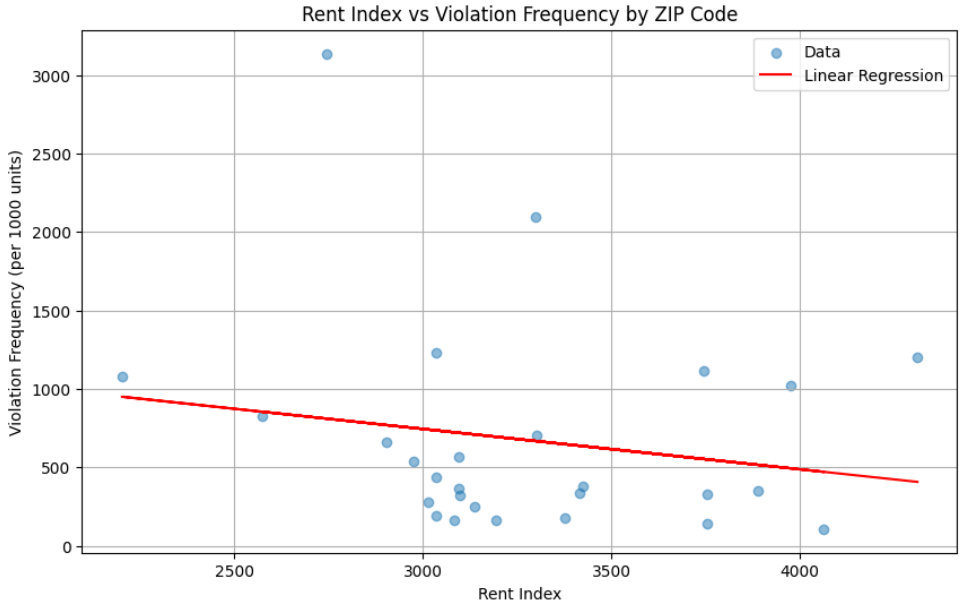

Figure 34: Rent Index vs Violation Frequency by ZIP Code

While the overall trend supports our hypothesis, the presence of outliers signifies the influence of additional factors. Some high-rent districts still experience a notable number of violations, indicating that high rent alone isn't a blanket deterrent to violations.

Next, we used the median household income data from incomebyzipcode.com to construct a socio-economic profile for each ZIP code. We then juxtapose this data with violation frequencies to investigate potential economic disparities.

The linear regression model depicted in the accompanying graph demonstrates a clear negative correlation. As median household incomes rise, the frequency of housing violations

per 1,000 units declines. The slope of our regression line is negative, and most data points fall close to this line, indicating a strong relationship between income levels and housing violation frequencies.

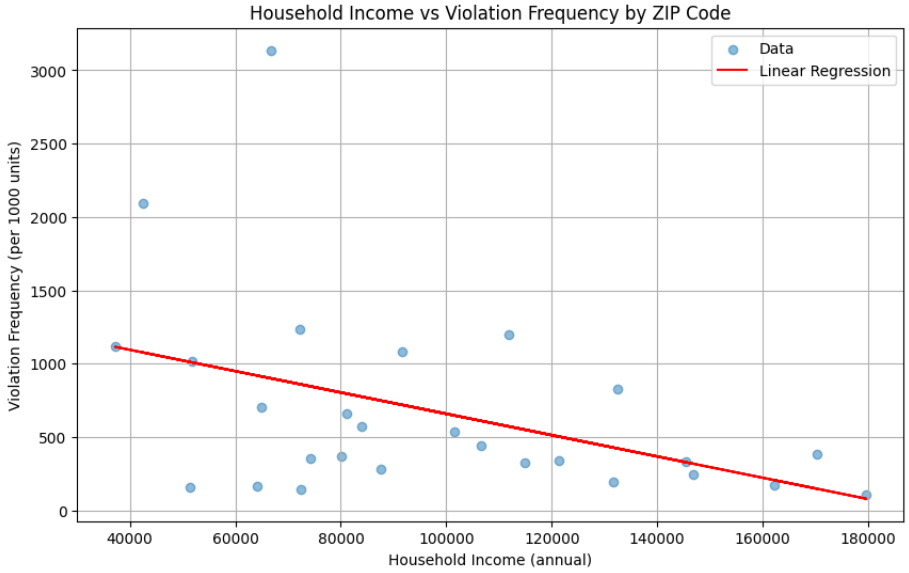

Figure 35: Household income vs Violation Frequency by ZIP Code

This inverse relationship may be indicative of higher-income households having more resources for property upkeep and adherence to housing regulations, thereby reducing the likelihood of incurring violations. Nonetheless, the trend's outliers suggest that income is not the sole determinant of housing conditions.

Delving deeper, beyond absolute income levels, we examined the proportion of income that households allocate towards rent, hypothesizing that a lower income-to-rent ratio may correspond to a lesser likelihood of housing violations, reflecting better living conditions and maintenance capacity.

Our regression model indicates a positive slope, suggesting that as the proportion of income spent on rent decreases, the frequency of violations per 1,000 units also diminishes.

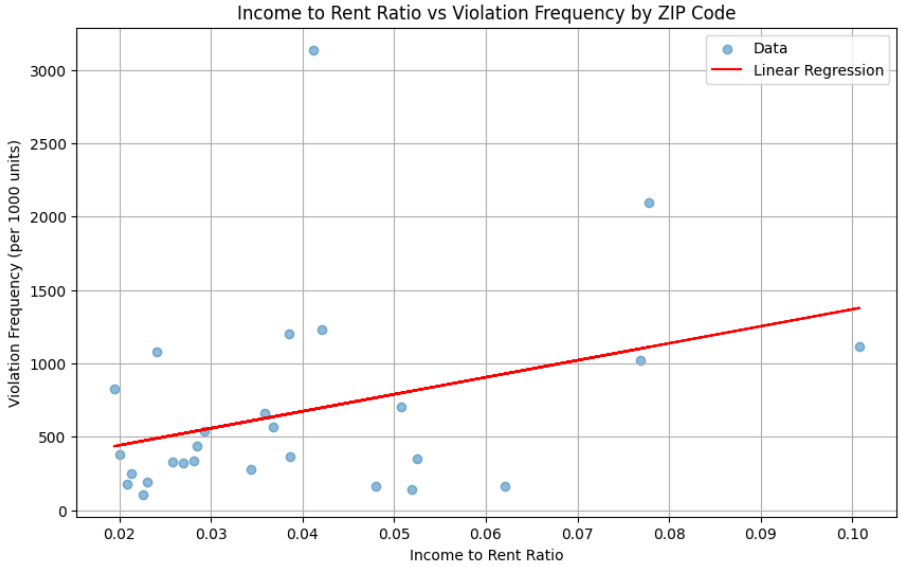

Figure 36: Income to Rent Ratio vs Violation Frequency by ZIP Code

This relationship, while present, is not as pronounced as the income-violation correlation, signaling that affordability plays a role, but its influence is subtler.

A lower income-to-rent ratio may indicate more disposable income for home maintenance and investment in living conditions, which could contribute to compliance with building standards. However, the modest slope of the trend line and the variability among data points remind us that affordability is one piece of a larger puzzle.

Conclusions:

Areas with higher Rent Index values, reflecting higher rents, generally report fewer violations per 1,000 units, suggesting better housing quality and maintenance.

Higher household incomes are associated with lower rates of violations, indicating that economic prosperity could translate into improved housing conditions and compliance.

The Income to Rent Ratio, while less pronounced, still underscores the importance of affordability and disposable income in the context of housing quality.

The data informs us that interventions aimed at improving housing conditions and reducing violations should be socio-economically sensitive, targeting not just the physical infrastructures but also the economic realities of residents. Policies could include targeted subsidies for housing maintenance in lower-income areas, incentives for landlords to improve property conditions, and community-based initiatives to raise awareness about housing standards.

4 Future Scope

While this project has provided valuable insights into building violations across Boston neighborhoods and their relationships with various demographic, environmental, and economic factors, there are several areas that could be explored further in future work:

1. Predictive modeling: The current analysis focused mainly on identifying correlations and trends. A logical next step would be to develop predictive models that could forecast future violation rates based on changes in key variables like population, education levels, rental prices, income, etc. This could help the city proactively target enforcement and allocate resources. Techniques like time series forecasting, regression modeling, and machine learning could be applied.
2. Causal analysis: This project uncovered interesting relationships, such as between education levels and violation rates. However, correlation does not imply causation. Future work could dive deeper into potential causal mechanisms behind these relationships through techniques like controlled experiments, difference-in-differences analysis, instrumental variables, etc. Understanding root causes is critical for designing effective interventions.
3. Cost-benefit analysis of interventions: Before implementing any new programs or policies to reduce violations, it would be prudent to conduct a rigorous cost-benefit analysis. This could involve estimating costs of increased inspections, educational campaigns, subsidies, etc. and comparing them to projected benefits like fewer violations, improved public safety, higher property values, etc. Advanced econometric modeling and sensitivity analysis could assist these evaluations.
4. Interactive dashboards: To make the insights from this analysis more accessible and actionable for stakeholders, interactive dashboards could be developed. These web-based tools could allow users to slice and dice the data by neighborhood, time period, violation type, etc. and visualize results. Real-time integration with the city's databases could keep the dashboards current. Collaborative features could also allow different departments to share knowledge and coordinate efforts more effectively.
5. Equity analysis: Future work should incorporate a deeper analysis of equity considerations. This could involve examining disparities in violation rates, enforcement actions, and resolution times across neighborhoods with different racial, ethnic and socioeconomic compositions. Spatial analysis could also reveal if adverse effects of violations are disproportionately impacting historically disadvantaged communities.

In conclusion, while this project has broken valuable new ground in understanding building violations in Boston, many productive avenues remain for further research and translating insights into impact. From predictive modeling to qualitative research to impact evaluation, a multifaceted approach can help ensure that every resident can enjoy safe, high-quality housing regardless of where they live in the city. Moreover, the analytical approaches pioneered here could potentially serve as a model for other cities grappling with similar challenges.

5 Individual Contribution

Youxuan Ma, Class of 2024, markma@bu.edu

1. Made mid-semester and final presentation slides. Delivered Early Insights presentation to the client and contributed to final presentation.
2. Organized and uploaded Early Insight Report materials to the repo and the folder online.
3. Helped with insights development.

Heyang Yu, Class of 2025, jhyyu@bu.edu

1. Solved the landlord key question.
2. Participated in the research on most affected neighborhood key question. Conducted the research on the extension of economic factors.
3. Made the early insights slide. Delivered midterm presentation to the clients and presented insight 8, 9 and 10 in the final presentation.

Jian Xie, Class of 2025, jianx@bu.edu

1. Engaged in analyzing base questions, offered insights and raised two extension proposals.
2. Assessed the quality of our answers at each stage, gathered feedback from team members after presentations, and synthesized solutions to improve our work.
3. Contributed to creating the slides, writing the final report, and delivering the concluding presentation.

Guanxi Li, Class of 2025, guanxili@bu.edu

1. Throughout the early insight, midterm, and final stages, coded the analysis process based on the ideas discussed by the team, programmed most of the code, and generated clean, usable new datasets and various types of visualizations.
2. Analyzed these results to assist team members who were responsible for creating the PowerPoint presentations and reports.
3. Was in charge of reporting on insight 7 in the final presentation.

Yuzhe Jiang, Class of 2025, jiangyz@bu.edu (Team Rep)

1. Proposed some optimization ideas for extension analysis, and make early insights and final reports.
2. Team representative, met with clients every week to promote project progress, arranged team development meetings and wrote weekly sprints and weekly scrum reports.
3. Introduced extension analysis part and presented insight 5&6 in the final presentation.