Boston Crime Data Analysis (2017-2022)

Individual Project COMP-3125

Isabella Crowe

***Abstract*—This paper analyzes the intersection between urban crime and student housing in Boston, Massachusetts. The study employs a data-driven approach, utilizing a publicly available dataset from Kaggle containing crime reports in Boston from 2017 to 2022. This paper thoroughly examines East Boston, Dorchester, Roxbury, and Charlestown while highlighting the Fenway districts. Crime trends were analyzed using visualizations such as pie charts and bar graphs, highlighting both the most common types of offenses and their geographic concentrations. Further, a spatial analysis was conducted using geographic bounding and interactive mapping with Python libraries, revealing a significant concentration of incidents in specific neighborhoods and streets. Results indicate that offenses such as larceny, vandalism, and motor vehicle accidents are most prevalent, with crime disproportionately concentrated in specific neighborhoods. The study highlights the trade-offs students face between affordability and safety when searching for off-campus housing in Boston, Massachusetts.**

***Index Terms*—Spatial Analysis, Python Analysis, Boston Neighborhoods, Crime Mapping**

I. Introduction

WHile this essay focuses on crime, it is part of a larger, more complex issue of the student housing crisis that occurs year after year in Boston and Cambridge, Massachusetts. A significant challenge for college students searching for housing is knowing where to begin. One major issue is that many colleges in Boston lack dedicated housing assistance centers for students looking to move off campus. Many students have encountered problems such as scams, inconsistent quality of listings, misleading apartment descriptions, and unverified information about landlords. Students are often willing to navigate these challenges, including false listings and inaccurate information, despite the risks to their safety.

Finding apartments in metropolitan areas can be tough, and Boston is one of the most expensive cities to live in. While an apartment's rent of $2,400 might seem high for the general public, it poses a significant challenge for college students. To adhere to the commonly recommended 30% of income for housing, a student would need to earn $8,000 monthly to afford to live in the metropolitan area comfortably. This is often unmanageable for most students, especially those balancing a full course load and extracurricular activities. Many students in the metropolitan area end up spending over 50% of their total income on rent, which constitutes a severe rent burden. This creates additional stress on top of the everyday pressures that college students face. [1]

As a result of these challenges, students often seek alternative housing options and are willing to make trade-offs, such as compromising on safety to find more affordable housing. This issue is particularly significant as many Boston students are unaware of where to find the best and safest housing options while searching for an apartment.[2]

In the past 25 years, Boston has been one of the safest megalopolises in the United States. Other cities, such as Memphis, Tennessee, and Los Angeles, California, have significantly higher crime rates compared to Boston. Memphis is similar in size to central Boston, while Los Angeles is comparable to greater Boston. In 2023, the nationwide rate of property crimes (which includes but is not limited to shoplifting, motor vehicle theft, embezzlement, burglary, and forgery) in the United States was 1,917 cases per 100,0000, equating to nearly 1.9%. However, in Boston, property crime was below the average, coming in at 1,820 cases per 100,000 or 1.8% of the population per year, lower than the national average.[3] Memphis (7.2%) and Los Angeles (2.5%) have higher property crime rates than Boston does.[4][5] Boston does have a higher-than-average violent crime rate. Violent crime encompasses offenses involving force or the threat of force, including murder, rape, robbery, and aggravated assault.[6] The FBI recorded 364 violent crimes per 100,000 in 2023. [7] Boston has nearly double the amount of violent crime rate than the nationwide average, coming in at 630 cases per 100,000. [3] Similarly, Memphis (2.4%) and Los Angeles (7.95%) have higher violent crime rates than Boston does. But these numbers mean little when one is looking for the safest housing within Boston. [4][5] To accurately analyze the crime rate in Boston, this paper will look into 13 different neighborhoods. This will include Allston/Brighton, Charlestown, Dorchester, Downtown, East Boston, Fenway/Lower Roxbury, Jamaica Plain, Mattapan, Roxbury, and West Robury/Roslindale.

II. Data sets

A. Source of Data set

This article will reference a data set from Kaggle about Boston Crime from 2017 to 2022. Kaggle is a platform for data scientists and machine learning practitioners created and monitored by Google LLC. Kaggle is a platform where individuals can publish datasets, build models collaboratively, and participate in competitions to solve data science challenges. Kaggle is widely regarded as a source of reliable datasets, especially in professional settings. Over the years, there has been suspicion that the datasets are too clean, lacking inconsistencies, duplicates, or missing values. [8]

B. Character of the Data sets

The dataset only contains crime-related records from Boston with specific attributes such as offense codes, locations of the crime, and dates. The CSV file on Boston Crime from 2017 - 2022 from Kaggle is divided into four categories:

Column Name, description, unit, and data type. The column name is the overarching nature of the crime (Ex. OFFENSE_DESCRIPTION could be burglary - commercial, towed motor vehicle, death investigation). The description is the specific events that happened during the arrest. The Unit is defined as the measurement unit for each column, if applicable. The measurements in this CSV file are either N/A, numeric, categorical, Date/Time, or geographic. The data type is the data stored in the columns, whether it is an integer, float, string, or datetime.

Each row represents an individual crime record identified by a unique Incident Number. The dataset includes both timestamps (OCCURRED_ON_DATE) and separate columns for YEAR, MONTH, DAY_OF_WEEK, and HOUR, facilitating time-based analysis. Other aspects of this dataset include geographical location as a tuple of latitude and longitude. There were a few missing values; some of the rows had 0 values for the latitude and longitude of missing offense codes. These incidents may not be tied to a specific location or may involve improperly recorded data. This dataset does not show evidence of merging multiple datasets. Each row corresponds to a separate incident record contained within the CSV file. The CSV file on Boston Crime from 2017 - 2022 from Kaggle has 31 rows representing individual incident reports and nearly 550 thousand columns.

III. Materials and Methods

The methods utilized throughout this project focus on analyzing and visualizing crime data in Boston, Massachusetts. The main goal of this analysis is to identify the most significant types of offenses and the most frequented areas where crimes occur, illustrated by pie charts, bar plots, and maps. The process involves reading data from the CSV file, filtering specific districts, cleaning the data, sorting the data, and sampling to create visual presentations of the plots and map of Boston’s crime. Cleaning the data is one of the most essential parts of this project, along with ensuring the data is accurate, reliable, and useful to the analysis. The data contains records specific to East Boston, Dorchester, Roxbury, and Charlestown, and the data is filtered based on the district. The dataset was reduced by systematically selecting every third row using iloc[::3]. This technique reduces the dataset’s size while maintaining a reasonably representative sample and is used throughout the entire project. Table I. shows the advantages and disadvantages of using pie charts and, for this project, includes overall data structure.

A. Pie Charts

While processing the data was that the Boston Crime from 2017 - 2022 CSV file for the Pie charts included ’DISTRICT’, ’OFFENSE_CODE_GROUP’. The unnecessary columns, ’SHOOTING’, ’Lat’, ’Long’, ’REPORTING_AREA’, ’UCR_PART’, ’INCIDENT_NUMBER’, ’OFFENSE_CODE’, ’DAY_OF_WEEK’, and ’HOUR’, were dropped using the drop() function to focus on the relevant columns. The pie charts were constructed using matplotlib, where the categories with percentages below 3% were hidden on the main pie chart. An additional legend was added to represent offenses that accounted for less than 3% of the total, with their corresponding percentage values displayed on the legend. This made the visualization of the pie chart easier to understand due to overlapping issues with labels when using matplotlib. The pie chart was generated using matplotlib’s plt.pie() function, with several customizations to improve clarity and relevance:

- The percentages of the top 20 offenses were calculated by dividing the count of each crime type by the total number of offenses, multiplying by 100 to convert to percentages.

$$Percentagei = (Counts/Counti)100$$

- The format_labels() function was defined to wrap offense labels exceeding 10 characters, adding line breaks to avoid overcrowding or overlapping text in the pie chart. This was incorporated into the code with a simple if statement and this formula: return .join(label.split())
- bbox_to_anchor was used to position the legend outside the chart, preventing overlap with the pie chart itself.

Advantages:	Disadvantages:
Simplification: By concentrating on the top 20 offensive categories, the visualization streamlines the overall data and offers clear insights.	Over-Simplification: Sampling every 20 incidents may include infrequent but significant offenses while potentially excluding other significant offenses.
Visualization: The pie chart visually presents the distribution of offense types, making the data easier to interpret at a glance.	Bias Within Data: Limiting the pie chart to only 20 categories may ignore the less common offenses, but getting a full picture of crime analysis is still important.
Filtering: The pie chart visually presents the distribution of offense types, making the data easier to interpret at a glance.	Loss of Data: Limiting the pie chart to only 20 categories may ignore the less common offenses, but it is still important to get a full picture of crime analysis.

TABLE I

B. Bar Charts

Bar charts were utilized to visualize crime data, focusing on the distribution of offenses by district. Two types of bar plots were created to explore analytical differences in crime in Boston.

1) District Chart: The first bar chart, Figure 5, illustrates the total number of crime incidents reported in each Boston police district from 2017 to 2022. This style of plotting was selected for its ability to compare crime volumes across different geographic regions visually. The assumption made is that each district functions independently with comparable reporting standards. Another is the total incident count, which reflects crime severity or frequency in each district. The advantages of bar charts of this nature are that they highlight regional crime concentration and are easy to interpret and compare.

The dataset is read by using Pandas with enforced data types (dtype=..) } to ensure uniform entries. Any non-relevant or missing information was removed, including rows with NaN values in the district-type column. Incident counts were calculated using value-count() and visualized with Seaborn's (a Python data visualization library) barplot() function. To enhance the interpretability of the bar chart, color intensity was mapped to the frequency of the crimes using Normalize( ), a function from matplotlib.colors. This function scales the count values between 0 and 1, allowing each bar color to correspond to its height. The normalized values are applied to a reversed Red-Yellow-Green ('RdYlGn_r') colormap to emphasize high-crime districts in red and low-crime districts in green, as illustrated in Figure III & IV: Dynamic enhancements were added to the count label, positioning it above each bar using the ax.text() method. For Example 2, the following code snippet demonstrates the process of getting a float between 0 and 1 to return an RGBA color for each bar in the plot:

$forpinax.patches$: (1)

$$p.set\_facecolor(map(norm(p.get\_height())))$$

The X-axis label is rotated 45 degrees to improve readability. Gridlines were added along the y-axis using plt.grid() The gridlines were made semi-transparent so they don't dominate the plot using the alpha parameter.

2) Street Crime Chart: The second bar chart, Figure III & IV, illustrates the total number of crimes on a street level. Similarly to Figure V, the street crime graphs also use the Pandas library, Matplotlib, and Seaborn libraries for data manipulation and visualization. To narrow the analysis of the district's street crime, a geographical bounding box was applied to the latitude (Lat) and longitude (Long) coordinates. The analysis was implemented using Boolean masking with empirically derived thresholds. For Example 2, the following code snippet demonstrates the process of boolean masking for Charlestown, Massachusetts:

$lat\_min = 42.3700$ (2)
$lat\_max = 42.3785$

$$(data['Lat'] \leq lat\_min)\&(data['Lat'] \geq lat\_max)$$

Subsequently, the dataset was aggregated at the street level. The value_counts() function was used on the STREET attribute to calculate the frequency of crime reports per street segment. This yielded a distribution of incident density across street locations within defined Charlestown subregions. This form of filtering isolates the incidents occurring within the approximate boundary of Main Streat in Charlestown, excluding adjacent areas such as the Navy Yard and Bunker Hill to enhance the geographic precision. The condition (data['Lat'] ¿= lat_min) ensures that the longitude values are greater than or equal to the minimum latitude (42.3700). While (data['Lat'] ¡= lat_max) ensures the values are less than or equal to the maximum latitude (42.3785). The second condition for longitude is an exact copy of example 1 however, one separates them with a & symbol and substitutes lat_min for long_min. This logical AND operation ensures that only the data points correspond to incidents within the specific geographic area. Columns deemed extraneous to the objective, such as UCR_PART, were not included in these plots. The same colormap was selected, seen in Figure III & IV, to highlight high-crime streets in red and low-crime streets in green.

C. Map

This section describes the technical methodology applied to the process of crime data by street and validates the results on an interactive map of Boston. The process of using Python libraries such as Pandas for data manipulation, Folium for mapping, and webbrowser for visualization creates the Map. The Map uses the columns Lat, Long, and STREET. The low_memory=False flag is specified to ensure efficient type inference during data loading when dealing with large datasets. Rows with missing values in the critical columns (Latitude, Longitude, and STREET) were removed using the dropna() function. The data was organized by counting the number of crime incidents on each street, using the 'groupby()' and 'size()' functions to calculate the crime count for each location. To visualize a map centered on Boston, the `folium.Map` function was implemented. `folium.Map` utilizes the coordinates for Boston, with a latitude of 42.3601 and a longitude of -71.0589, set at a zoom level of 12 for an optimal view of the city. A color gradient was applied to the markers based on crime counts, ranging from dark red for streets with over 5,000 crimes to green for those with fewer crimes. To enhance the map's usability, a legend and title were included using HTML structure.

IV. RESULTS

A. Pie Charts

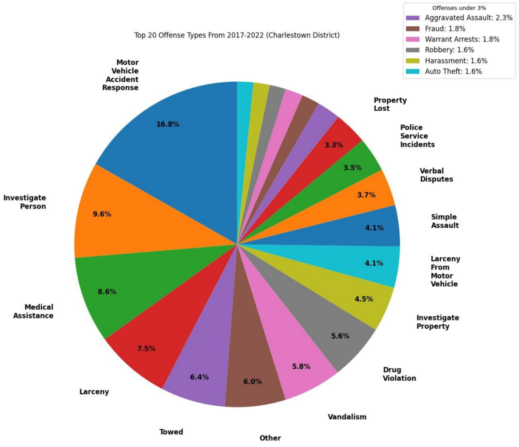

Fig. 1. Top Offense Type: Charlestown

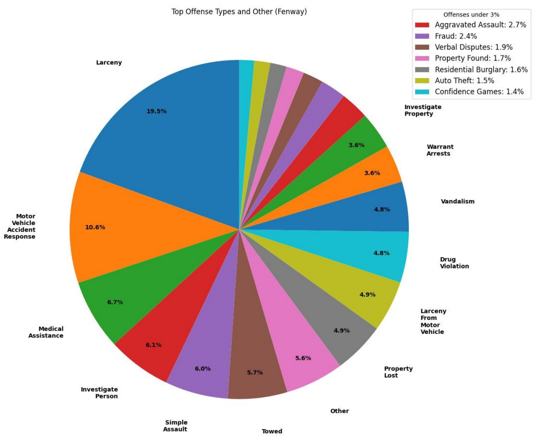

Fig. 2. Top Offense Type: Fenway

The Pie chart utilized throughout this project focuses on analyzing and visualizing crime data in Boston, Massachusetts. These charts specifically look into the towns in Boston with the most crime and least amount of crime. The pie charts specifically look into the surrounding towns of East Boston, Dorchester, Roxbury, Charlestown, and Fenway. The charts show the top 20 offense categories with an additional legend showing only the crimes under 3% shown in Figure I II of five (Dorchester, Roxbury and East Boston not shown).

Figures I & II present the distribution of crime types in the data set by district expressed as a percentage of the total reported incidents in Charlestown, Dorchester, East Boston, Fenway, and Roxbury. The most frequently reported offense was larceny, accounting for 9.6% of all incidents in these five districts. This was followed by vandalism at 5.78%, simple assault at 5.52%, and drug violations at 5.52%. These percentages were calculated using the formula $Percentage = (Counts/Count)100$, where 'Counts' refers to the number of offenses of a specific type, and 'Count' represents the number of districts—5 in this case. Notable Motor Vehicle accidents also appeared frequently, with district-level rates including Charlestown at 16.8%, East Boston at 14.6%, Dorchester at 14.5%, Roxbury at 13.8%, and Fenway at 10.2%. Motor Vehicle accidents account for 13% of all police activity in Boston. These values indicate that traffic-related incidents represent a considerable share of reported activity in districts.

Across the five districts, minor offenses exhibit a consistent pattern of low but notable occurrences. Offenses such as Fraud, Auto Theft, Warrant Arrest, Aggravated assault, and Property loss are recurrent across all five districts, each typically ranging between 1% and 4%. One average fraud appears to account for 1.7% of all incidents, Auto Theft at 2.2% Warrant Arrests at 3.1%, Aggravated Assault at 2.9%, and Property Loss at 3.1%. While these percentages individually are considered low, their cumulative presence represents a non-negotiable portion of the total reported offenses.

B. Bar Plot

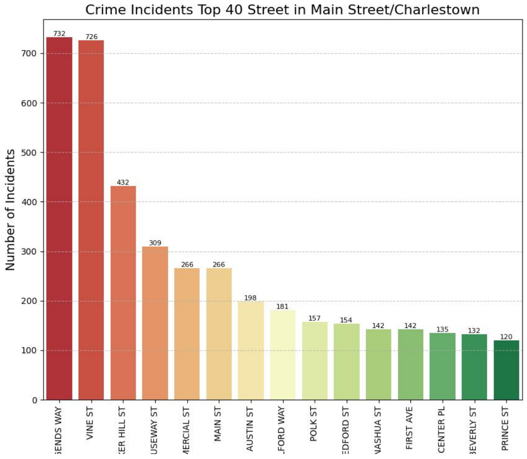

Fig. 3. Top 15 Street Crime Level: Charlestown

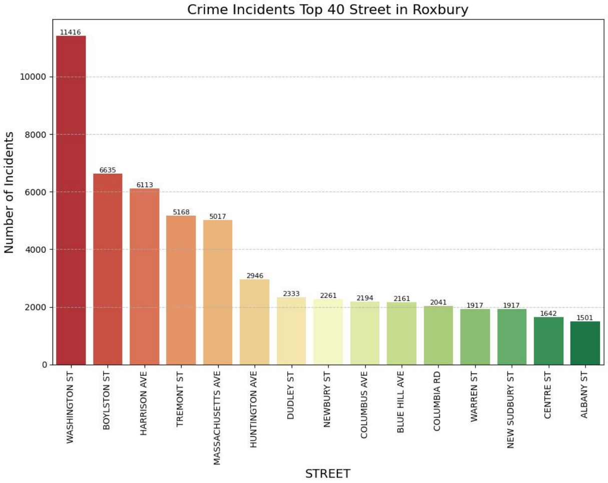

Fig. 4. Top 15 Street Crime Level: Roxbury

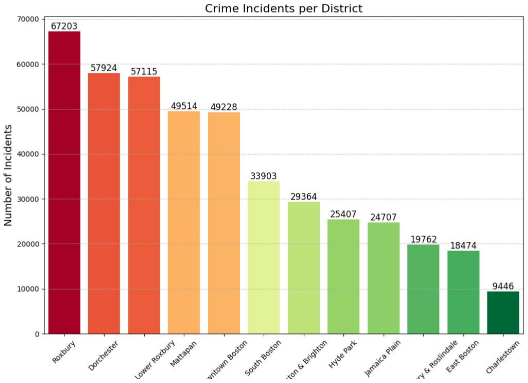

Fig. 5. District Crime level

1) Street Crime Chart: The analysis of crime incidents across four neighborhoods—Fort Hill, Main

Street/Charlestown, Navy Yard, and Roxbury—reveals significant disparities in the distribution and intensity of crime across the top 50 streets in each area.

Roxbury exhibits the highest crime incidents overall in this analysis, with Washington Street recording a staggering 11,416 incidents, far surpassing all other streets across neighborhoods. Boylston Street (6,635) and Harrison Avenue (6,112) also report elevated crime levels, creating a steep drop-off from the highest-ranked streets to those further down the list. The top 10 streets in Roxbury all exceed 2,000 incidents, ranging from motor vehicle accidents and larceny to simple assault and drug violations. This widespread crime intensity rivals and exceeds that of all districts in all of greater Boston. Fort Hill, a neighborhood and historic district in Roxbury, experiences higher crime rates than other Boston sub-districts. Centre Street has the most crimes with 839 incidents, followed by Columbus Ave Street (580 incidents). Crimes here are still significant but are more evenly spread across the top 50 streets, each registering over 100 incidents.

Charlestown remains the area with the lowest reported crimes in the greater Boston area. Legends Way has the highest amount of crime, with 732 incidents, and the remaining streets have a narrow and relatively low spread within 50th ranked street Hancock Street showing 2 incidents over 4 years. Navy Yard presents a more moderate crime profile, with the lowest reported crimes in this study. Bunker Hill, the highest-ranked, had only 361 incidents, and the remaining streets showed less than 100 incidents in the past 4 years.

An analysis of commercial landscapes in Charlestown and Roxbury reveals a distinct difference in retail composition, which may partially account for the variations in the crime incidents observed across neighborhoods. Charlestown features a predominance of lifestyle-oriented stores and high-end grocery chains such as Whole Foods. These establishments generally attract more affluent customers and are associated with well-maintained infrastructure, private security, and limited operation hours. Such characteristics contribute to a possible reduced incidence of opportunistic or violent crimes due to minimized cash-based transactions and the promotion of surveillance.

Conversely, Roxbury exhibits a broad spectrum of community-focused retail, including convenience stores and service-based establishments concentrated around high-traffic transit corridors such as Nubian Square. These Businesses often operate during extended hours and may rely heavily on cash transactions, increasing vulnerability to robbing and petty theft. Furthermore, Limited security infrastructure and lower commercial investment may reduce deterrents to criminal behavior. This contest is consistent with the crime incident distribution shown in Figure III & IV, where the major streets, such as Washington St and Boylston St, report significantly higher incident counts compared to Charlestown and Fort Hill.

2) District Crime: The analysis of incident data across Boston neighborhoods reveals substantial disparities in reported occurrences. Roxbury recorded the highest number of total incidents. totaling over 67 thousand, followed by Dorchester (57,924) and the South End (57,115). These three neighborhoods alone account for a significant proportion of the total crime analyzed in the greater Boston area. Mattapan and Downtown Boston have nearly identical incident reports with 49,514 and 49,228 incidents, respectively. Brighton (33,902), Hyde Park (29,364), East Boston (25,407), and Jamaica Plain (24,707) would be concerned with the mid-range incident count. Lower Levels of reported incidents were found in West Roxbury (19,762), Roslindale (18,474), and Charlestown with just under 9,500. Overall, the data suggests a geographic concentration of incidents in Roxbury, Dorchester, and Southend, which consistently exhibit the highest frequencies.

C. Map

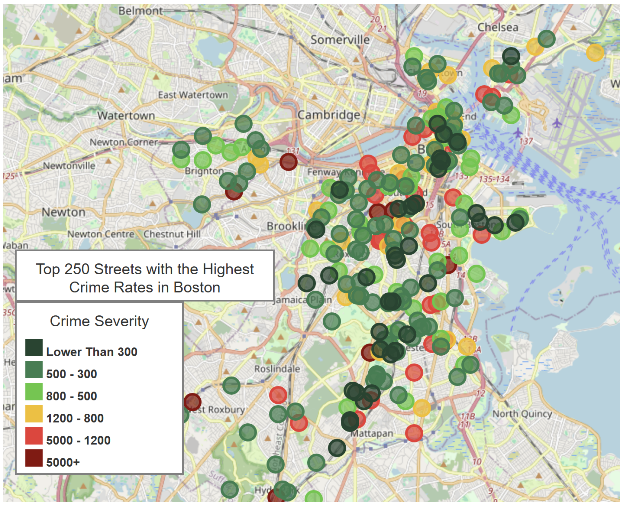

Fig. 6. Figure 3. Charlestown Offense Types (Reused)

To visualize the spatial concentration of crime across Boston, an interactive Choropleth-style map was generated using Folium, identifying the top 250 streets with the highest crime count from 2017 to 2022. The map plots each street as a circle marker with color intensity indicating the severity of total reported incidents. Darker red tones signify streets exceeding 5,000 incidents, while dark green denotes areas with fewer than 300 cases.

The spatial pattern shows a high concentration of crime in central and southern Boston, particularly in neighborhoods like Roxbury, Dorchester, and parts of the South End and Brighton. These regions exhibit multiple hotpots with incident counts surpassing 1,200, with a subset exceeding 5,000. Washington St in Brighton has an extremely high crime rate with over 20 thousand incidents, and Blue Hill Ave next to the Franklin Park Zoo has an incident count of 10,358. Tremont St. and Massachusetts Ave. in the south end both have incident counts of over 6,000. In contrast, areas like Charlestown and East Boston show significantly fewer high-severity markers, suggesting lower overall crime prevalence in those districts. This visualization not only confirms findings from prior bar plot analyses but also emphasizes the spatial clustering of street-level crime. The heat pockets correlate strongly with major arterial streets and commercial corridors, such as Washington St. and Blue Hill Ave. This could align with socio-economic and urban disparities.

V. Discussion

When building such a large project in little time, there are always shortcomings and places where the developer wishes they had focused more intently on specific areas. The primary limitation of this project lies in the constrained scope of analysis conducted on the CSV file that was examined. This restriction may prevent a comprehensive understanding of the data and its implications, as a broader approach could lessen the depth of the project when there are time restraints.

This project needed to have more raw data analyzed rather than making maps and plots that were analyzed the same way. This data could show how many crimes were committed throughout the day by the district in Boston, which would have made a strong addition to the analytical side of the project. One could then make a visual representation, animating throughout the day to show a user when the most dangerous times to do regular activities are and when it is the safest. Another area that this project could improve upon is the type of offense that happens the most on specific roads. The last improvement would be using the Boston PD crime data rather than Kaggle's. Even though Kaggle is a reputable source, getting data from the origin is always the most factually accurate.

VI. Conclusion

This paper presents a comprehensive study of crime patterns in Boston through the lens of student housing challenges, illuminating the often-overlooked relationship between affordability, safety, and informed decision-making. At the core of the issues lies the severe housing burden many students endure throughout their college experience. With rent prices racing to unaffordable levels, students are forced to compromise on important housing factors, including location, quality of living arrangements, and most crucially, safety. While Boston as a whole reports that property crime rates lower than the national average, its violent crime rate is significantly higher. These broader statistics fail to reflect the vastly different safety profiles of Boston's diverse neighborhoods. By separating crime data across key districts, highlighting Roxbury, Dorchester, East Boston, Charlestown, and Fenway, this paper identifies which communities pose a greater risk and which may offer safer, more secure housing options for students.

Data analysis and visualization, Pie charts, Bar plots, and an interactive map provide insights from the data set. Pie charts offered a district-level view of the most common offenses, such as larceny, vandalism, and drug violation, while highlighting patterns of minor but frequent crimes. Bar Barts dove deeper into crime frequency by both district and street, revealing a concentration of incidents in neighborhoods like Roxbury and Dorchester while Charlestown, although geographically smaller, has a significantly lower overall crime association than its neighboring districts. The study showed that crime intensity is not evenly distributed across Boston. Some neighborhoods, such as Roxbury and Dorchester show consistently higher crime rates, particularly along specific corridors (Washington Street and Harrison Avenue), while other districts report relatively fewer offenses. These findings are critical for students attempting to make housing choices but may lack access to easy-to-interpret safety data.

By focusing on just five Boston districts and sampling every third row in the dataset may have resulted in the loss of some rare events. Additionally, even though pie charts and box plots paint a strong picture of data analysis, they often obscure less frequent but impactful crimes such as aggravated assault or homicides. Future research could benefit from incorporating severity-weighted crimes, which would offer a more balanced risk assessment for student tenants.

In conclusion, the convergence of crime data and student housing in Boston reveals a possible linked where students must frequently choose between affordable housing and personal safety. The data analysis in this paper provides resources to help students overcome the challenges of finding housing while prioritizing security and accessibility.

VII. Acknowledgment

Thank you, Assistant Professor Weijie Pang and Annalise Cowing

References

[1] I. Crowe, "The student housing crisis in boston," https://medium.com/the-student-housing-crisis-in-boston/the-student-housing-crisis-in-boston-f928aa0adafc, 2022, accessed: Apr. 12, 2025.

[2] uAspire, "Unseen Issue for Higher Education in Massachusetts: Housing Insecurity Intensified," https://www.uaspire.org/news-events/unseen-issue-for-higher-education-in-massachusetts-housing-insecurity-intensified, 2021, accessed: Apr. 12, 2025.

[3] @usnews, "How safe is boston, ma?" https://realestate.usnews.com/places/massachusetts/boston/crime, 2022, accessed: Apr. 12, 2025.

[4] ——, "How safe is memphis, tn?" https://realestate.usnews.com/places/tennessee/memphis/crime, 2022.

[5] ——, "How safe is los angeles, ca?" https://realestate.usnews.com/places/california/los-angeles/crime, 2022.

[6] L. Gilder, "What does new fbi data show about us violent crime?" *BBC.com*, Sep. 2024. [Online]. Available: https://www.bbc.com/news/articles/c4glxxreed7o

[7] Federal Bureau of Investigation, "Fbi releases 2023 crime in the nation statistics," https://www.fbi.gov/news/press-releases/fbi-releases-2023-crime-in-the-nation-statistics, Sep. 2024.

[8] S. Negi, "Boston crime dataset (2017 - 2022)," https://www.kaggle.com/datasets/shivamnegi1993/boston-crime-dataset-2022?select=csv_2017_2022.csv, 2022, accessed: Apr. 12, 2025.