**Crime Rates and Housing Vacancies: A State-Level Study**

Group 7: Alyssa Day, Isabella Kleckner, Samy Babikerali, Togaa Lavien

The University of North Carolina at Charlotte

DTSC 3602

Dr. Nadia Najjar, Dr. Shannon Reid

**Contents**

Introduction ……………………………………………………………………………………… 3

Literature Review & Background ..……………………………………………………………… 5

Dataset Description ..…………………………………………………………………………….. 9

Methodology …………………………………………………………………………………….15

Analysis ………………………………………………………………………………………….19

Discussion………………………………………………………………………..………………30

Conclusion …..……………………………………………………………….………………… 32

References ……………………………………………………………………………………… 33

**Introduction**

In Detroit, Michigan, over 20,000 properties remain vacant, leaving neighborhoods to confront a harsh reality. Each additional abandoned home raises the likelihood of nearby property crime by 19% (Detroit Land Bank Authority, 2021; FBI, 2024). Vacant properties are not just remnants of economic downturns; they are hubs for illegal activity and symbols of urban decline. Research has shown that vacant properties contribute significantly to social and economic instability, leading to further disinvestment in affected neighborhoods (Chen & Rafail, 2020). Across the nation, this pattern repeats itself. While national vacancy rates have remained stable at 7.4%, the accompanying surge in property crime—up by 10.5% since 2022—illustrates a growing divergence between physical abandonment and social consequences. This divergence underscores the urgency of understanding how vacancy and crime are interconnected in different housing markets, especially as such trends amplify urban decay. This rise in property crime not only destabilizes communities but also places immense financial pressures on public systems, as evidenced by the $282.1 billion allocated to law enforcement and corrections in 2021 (BJS, 2022). Declining property values in affected neighborhoods worsen the situation, further eroding economic stability and perpetuating cycles of poverty (Whitaker & Fitzpatrick, 2012). These interrelated challenges—rising crime, public expenditures, and economic instability—underscore the urgent need to explore innovative solutions that address the root causes of these issues.

This study investigates the question: **How do increased non-violent crime levels predict residential vacancy rates in local housing markets across different states?** Unlike conventional reactive methods, which often focus on law enforcement or post-abandonment interventions, this study combines established frameworks with predictive modeling to offer a more proactive approach. Moving beyond a purely theoretical analysis, this study examines how non-violent crimes, such as property offenses and drug-related activities, destabilize housing markets. Social Disorganization and Broken Windows theories provide the theoretical foundation for understanding how economic instability and visible signs of decay drive both property abandonment and crime (Sampson & Raudenbush, 2004; Wilson & Kelling, 1982), ultimately informing intervention strategies. By combining these frameworks with predictive modeling, this research provides actionable insights into the relationship between crime and housing instability.

By restoring vacant properties and strategically targeting high-crime areas with data-driven interventions, communities can achieve transformative outcomes: safer neighborhoods, revived housing markets, and enhanced public revenues. These interventions could mitigate the negative economic and social impacts of vacancies, offering sustainable solutions to urban decay. These strategies stand apart from traditional enforcement-based solutions by prioritizing preemptive resource allocation and economic incentives. For instance, tax abatements for private redevelopers encourage investment in distressed properties, while grant-funded revitalization projects build social cohesion. Predictive modeling identifies high-risk areas, ensuring resource allocation is both efficient and impactful. By integrating these data-driven approaches with urban policy frameworks, this research demonstrates how proactive measures like predictive modeling can reshape public strategies for addressing urban decay and housing market volatility.

To ensure the thoroughness of our research, we adopt a multifaceted approach, combining theoretical insights with extensive data analysis. Using predictive analytics, this study identifies actionable patterns, focusing on key predictors of vacancy-induced crime while streamlining the analysis of state-level trends. By leveraging diverse modeling techniques, advanced statistical tools, and geospatial visualizations, this study uncovers nuanced relationships between crime and vacancy rates across various states. This research will contribute to the growing body of work on urban revitalization by offering new insights on how predictive modeling can guide targeted intervention. This approach allows for robust comparisons and highlights key predictors of vacancy-induced crime, offering practical solutions for urban revitalization efforts.

Ultimately, this research outlines a framework for addressing urban decay by emphasizing strategies such as funding property renewal initiatives and targeting high-risk areas for intervention. By focusing on proactive measures, this research seeks to break the cycle of decline and foster long-term economic stability. Collaborative efforts from government, businesses, and residents will play a key role in shaping sustainable urban resilience.

**Background**

The relationship between crime and residential vacancies is well-documented, particularly in urban studies and criminology. However, existing literature often neglects the interplay between state-level policies and localized factors such as social cohesion and visible disorder. This gap highlights the need to explore how systemic interventions, specifically at the state level, influence neighborhood decline and vacancy cycles. By addressing this gap, this study aims to investigate how state-level crime rates, policies, and housing stability influence vacancy rates across different regions.

Social Disorganization Theory and Broken Windows Theory provide a complementary framework for understanding neighborhood decline. The former explains how weakened social cohesion fosters crime, while the latter highlights how visible neglect, such as abandoned properties, exacerbates disorder (Shaw & McKay, 1942; Wilson & Kelling, 1982). Despite their useful insights, these theories frequently disregard deeper systemic connections, such as the role of state policies. This study aims to build upon these theories by incorporating the influence of state-level crime policies and housing protections, offering a more comprehensive view of how such interventions contribute to neighborhood stability and residential vacancy trends.

For example, tenant protection laws in California reduce housing turnover and stabilize neighborhoods, contrasting with states where weaker protections exacerbate residential instability. This variance in state-level policies is critical to understanding how residential vacancies differ between states, as it provides insights into the role state policies play in shaping neighborhood dynamics. However, these policy impacts are rarely analyzed in conjunction with localized factors like crime rates or visible neglect. By addressing these linkages, this research bridges theoretical and practical perspectives, offering a comprehensive model of neighborhood decline. While these theories provide a broad understanding of neighborhood decline, applying them to real-world examples, such as high residential turnover and economic disadvantage in cities like Detroit, reveals the nuanced interplay of localized and systemic factors driving these trends.

***The Relationship Between Crime Rates and Housing Vacancies***

Neighborhoods with low social cohesion, high residential turnover, and economic disadvantage often lack the informal controls needed to mitigate crime. These dynamics lead to residential instability and amplify property vacancies as residents leave areas perceived as unsafe (Sampson & Groves, 1989). In addition to crime, economic factors such as job loss, housing affordability, and broader market shifts contribute to vacancies. At the state level, variations in crime rates can exacerbate these issues, with higher crime states like Michigan experiencing greater vacancy rates in urban areas such as Detroit. Detroit exemplifies how systemic housing shortages and high eviction rates displace residents, contributing to cycles of decline where vacant properties attract criminal activity, such as vandalism or drug offenses (Porter et al., 2011). Understanding these dynamics highlights the need to integrate macro-level policies with localized interventions, a gap this study addresses.

***Theoretical Frameworks: Crime, Vacancy, and Community Decline***

The relationship between crime and residential vacancies is often rooted in two foundational theories: Social Disorganization Theory and Broken Windows Theory. Social Disorganization Theory highlights how weakened social networks and informal controls foster community instability, leading to heightened crime rates and residential vacancies (Shaw & McKay, 1942; Samspon, 2012). For instance, after Hurricane Katrina, abandoned houses in New Orleans became hubs for criminal activity, illustrating how a lack of social cohesion and stability can exacerbate cycles of decline. Similarly, neighborhoods like Detroit, marked by systemic economic challenges, reveal how social disorganization amplifies vulnerabilities in already struggling communities (Porter et al., 2011). By considering state-level variations in crime policies and housing protections, this research aims to explore how these factors shape the neighborhood dynamics described by these theories.

In contrast, Broken Windows Theory emphasizes the impact of physical neglect—such as vacant homes and graffiti—in signaling a lack of community investment and fostering environments conducive to crime (Wilson & Kelling, 1982). This was evident in Chicago, where the targeted demolition of high-rise public housing reduced crime significantly by addressing visible signs of disorder (Van Doren, 2016). Richmond, Virginia’s Operation Peacemaker Fellowship further illustrates how addressing both structural and social disorder can stabilize high-crime neighborhoods and break cycles of vacancy and criminality (Braga et al., 2019). State-level policies that address visible neglect and structural decay have a critical role to play in breaking these cycles and promoting neighborhood stability, as this study will examine.

Together, these theories suggest that both physical and social factors contribute to neighborhood decline, and addressing them can prevent cycles of vacancy and crime. This study seeks to expand on these frameworks by incorporating the effects of state-level crime rates and housing policies, offering a more nuanced understanding of how these factors shape the persistence of residential vacancies.

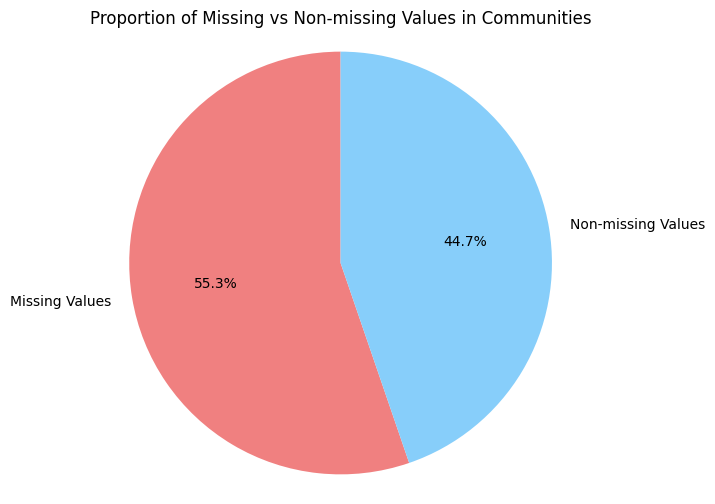
***Broader State-Level Implications and Policy Considerations***

At the state-level, policies such as housing regulations and crime prevention initiatives shape neighborhood-level outcomes by addressing structural and visible factors that contribute to vacancies. Eviction laws, for example, impact housing turnover; states with stronger tenant protections, like California, tend to see greater stability, whereas weaker protections result in higher turnover and vacancies. This aligns with Social Disorganization Theory, where stable communities are less vulnerable to decline and crime. Similarly, policies addressing visible neglect, like the Low-Income Housing Tax Credit (LIHTC), which revitalizes high-vacancy areas, support the Broken Windows approach by improving structural stability and reducing disorder. These state-level interventions are crucial in shaping the dynamics between crime and vacancy rates, which is the focus of this study.

However, the clustering of social housing remains a critical challenge. Research shows that such concentration can exacerbate crime by increasing proximity to offenders (Farrall et al., 2015). Policymakers should consider dispersing social housing and promoting mixed-income developments to balance crime prevention with long-term social equity goals, ultimately fostering resilient communities (Chaskin & Joseph, 2014). This study will examine how state policies regarding social housing density and distribution can influence crime and vacancy rates at the state level. By integrating these policies with insights from the theories discussed, this research aims to demonstrate how state-level interventions can effectively break the cycles of vacancy and crime.

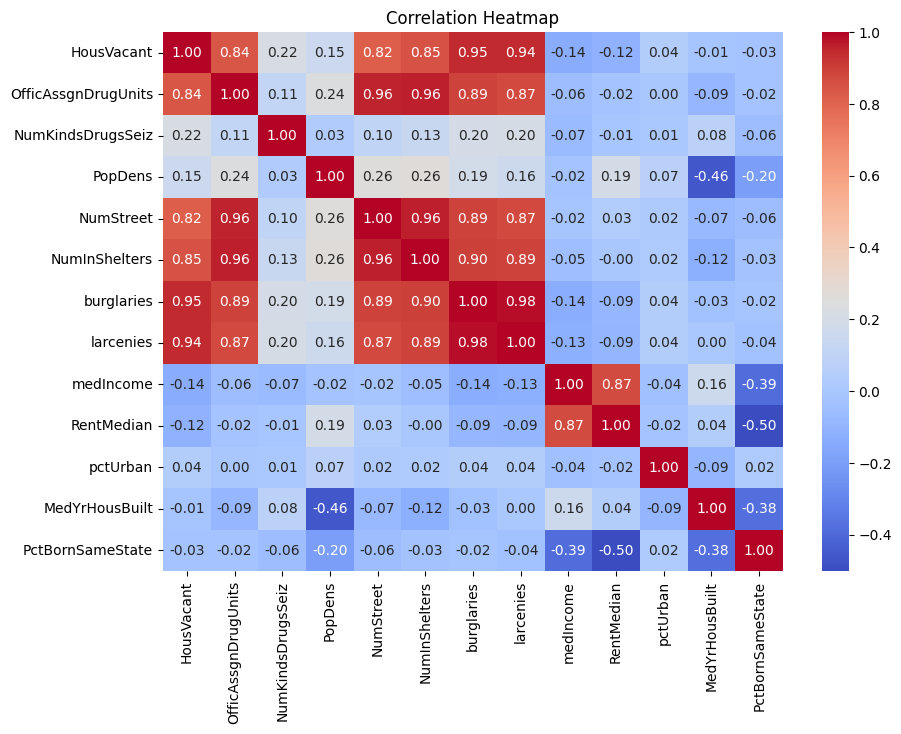
**Dataset Description**

For this research, a dataset combined from the US Census, LEMAS survey, FBI UCR was used. The dataset, known as “Communities and Crime”, includes socio-economic, law enforcement, and crime data from the early 90s. The dataset was preprocessed by undergoing a comprehensive normalization process, which provides advantages and drawbacks. The integration of the dataset provided the core attributes for a research between communities and crime to be conducted. However, it also caused an increased number of missing values across certain states. Additionally, the normalization process wasn’t able to preserve extreme values, but the relationship between attributes, provided a standardization of attributes in relation to their respective populations, which makes the attributes comparable across different scales (population).



*Figure [1] - Proportion of missing values in Communities*

Communities and Crime contained many missing values, specifically from the data integrated from the LEMAS survey. The main attribute of missing data in the LEMAS survey were the communities, with over 55% of the records missing. Consequently, The research conducted a state level analysis over a deeper micro level examination, to ensure a more accurate measure of attributes affected by the missingness of communities.



*Figure [2] - Correlation Heatmap of all relevant variables*

The key variables regarding the research question are extracted from the dataset by utilizing a correlation matrix. The independent variables (predictors) that directly contributed to our research question are location, non-violent crime, no. of assigned drug units, no.of (homeless) shelters, rent, house price, and median income. While house vacancy is the dependent variable (predicted).

As this research was conducted at a state level, all the records of the Communities and Crime dataset were grouped by state. This was done by aggregating the sum of all the communities in each state. Crime was measured by larcenies and burglaries due to the fact they were pertinent to the most critical variable of this research, house vacancy; their pertinence was determined by their correlation with house vacancy where both variables had a correlation over 94%.

Another excellent indicator of criminal activity is the variable, number of assigned drug units; which is the sum of all assigned police drug units in each state; this variable had a correlation of 84% with house vacancy.

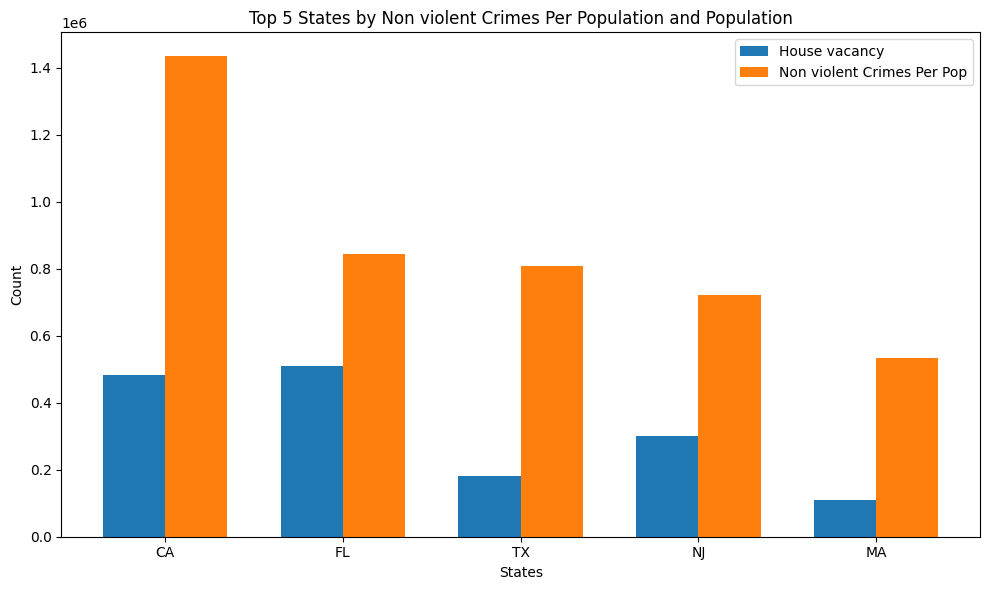
While there were other variables to measure homelessness, the number of homeless shelters (NoHS) was the greatest measure of homelessness at a state level; the correlation between NoHS and house vacancy is 85%.

Lastly, one of the primary subjects of this research was to verify the relation between the housing market and house vacancies. The housing market was measured by 2 key variables, median rent, and median house prices.

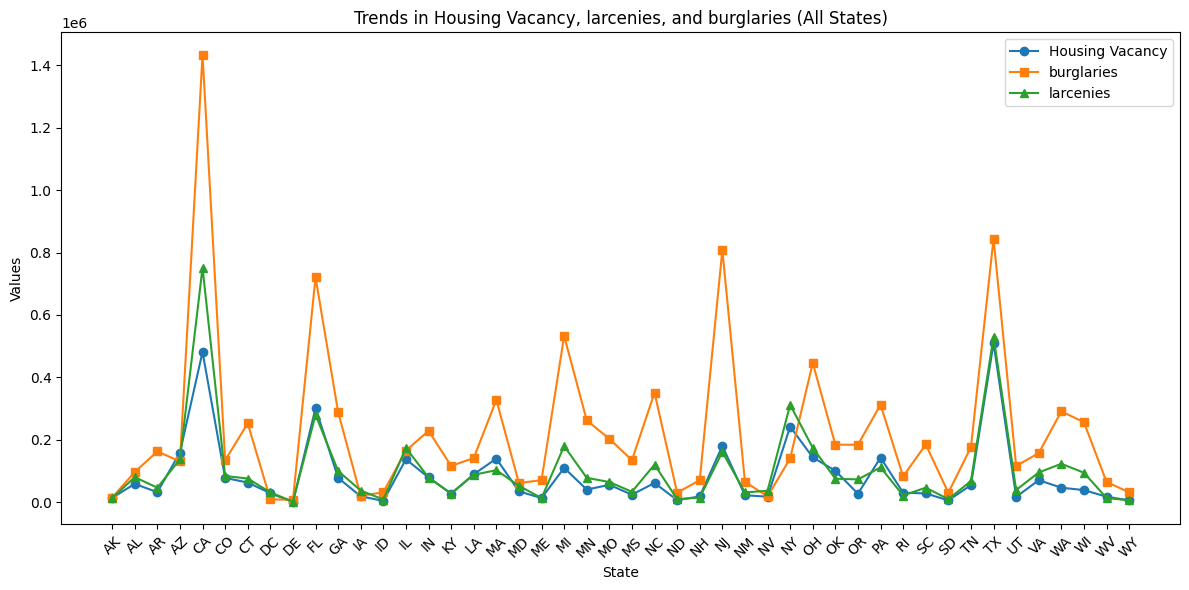
*Table [1] - Highest correlated features with house vacancy*

| **Subject** | **Measure (per state)** | **Correlation with House Vacancy** |
| --- | --- | --- |
| Crime | Burglaries | 0.945 |
|  | Larcenies | 0.944 |
|  | No. of Assigned Drug Units | 0.836 |
| Homelessness | No. of Homeless in Shelters | 0.849 |
| Housing Market | Median Rent | 0.118 |
|  | Median Income | 0.140 |

Additionally, some supporting variables on the Communities and Crime dataset were the number of different types of drugs seized, which measures the diversity of drug distribution and drug related crimes; and population per state, which had an exceptionally high correlation between multiple key variables used in this research.

*Figure [3] - Top 5 States by Non-Violent Crimes Per Population*

Further analysis of the data revealed significant trends between rising non-violent crime rates and housing vacancies across various states. To illustrate this relationship further, **Figure [3]** illustrates a comparison of housing vacancies and non-violent crimes per population in the top five states with the highest crime rates. This visualization highlights the potential relationship between increased crime and higher vacancy rates. .



*Figure [4] - Trends in Housing Vacancy, larcenies, and burglaries*

To further break it down, **Figure [4]** illustrates the trends of house vacancy, larcenies, and burglaries across every state, again excluding Kansas and Vermont. This chart highlights a very strong reaction between housing vacancies and larcenies, while the relationship with burglaries appears on a different scale. This is an unexpected pattern as burglaries typically involve break-ins, which would imply they should be more closely related to housing vacancies, but that wasn’t the case.

The integration of the dataset provided the core attributes for research between communities and crime to be conducted, however, it also caused an increased number of missing values across certain states. The normalization process wasn’t able to preserve extreme values, but the relationship between attributes provided a standardization of attributes in relation to their respective populations, which makes the attributes comparable across different scales (population). However, not all variables were standardized, additional variables like ‘OfficeAssignDrugUnits’ need to be standardized against population to ensure a consistent scale across this research. More feature engineering steps are also needed to prepare the dataset such as managing outliers, and addressing missing values.

**Methodology**

To investigate the length to which increased levels of crime contributed to residential vacancies, this research utilized Python with a variety of libraries. The research starts by defining and selecting the dependent and independent variables. While different sets of Independent variables were used, House Vacancies remained unchanged as the dependent variable throughout this research. This step is then followed by Feature Engineering, which addresses missing values, outliers, and variables; Baseline models, both categorical, and continuous; Advanced/Complex models training; Cross validation of the model; and finally, tool selection.

The feature engineering process involved several key steps to prepare the data for modeling. First, we removed non-predictive columns such as communityName, state, countyCode, communityCode, and fold using the drop method, as these identifiers do not contribute to predicting the target variable. Next, missing values in the dataset were handled using a KNN Imputation technique. Specifically, we replaced any placeholder values ('?') with 0 and applied the K-Nearest Neighbors imputation with n\_neighbors=2 to estimate missing entries based on the nearest available data points.

Outliers were handled using the interquartile range (IQR) method, flagging extreme values in variables like ‘HousVacant’, ‘larcenies’, and ‘burglaries’. This step made it possible to manage outliers carefully, which is important for achieving accurate analysis.

Additionally, a new column was introduced, vacancy\_class, which is based on the HousVacant variable to categorize the level of housing vacancy. As part of preprocessing, HousVacant was classified into three distinct categories - “Low”, “Medium”, and “High” based on defined thresholds. This approach provided a more structured framework for analyzing and predicting vacancy levels. The distribution of these categories was examined using the value\_counts method to assess the balance across the different classes. Additionally, the variable OfficAssgnDrugUnits was standardized by taking its ratio against population density for every community.

In this research 2 baseline models were created as a reference to justify or refute the development and use of more complex models. The first baseline model is used as a reference for the linear regression model. This baseline model predicted all HousVacant observations against its mean value; the 2 metrics used to evaluate this model are root mean square error, and mean absolute error. The second baseline model is used as a reference for the remaining categorical models (Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes). The baseline model predicts vacancy\_class against its mode. The mode was extracted from 2 different bins of vacancy\_class. The first class was distributed by using the lower (Q1) and upper (Q3) quartiles, where all observations above Q3 were marked as High; all observations Below Q1 were marked as Low; and all other observations as Medium. The second bin was the balanced vacancy\_class mentioned in Feature Engineering. This model was evaluated using accuracy.

To begin the feature selection process, a linear regression model was created using only variables related to crime to predict house vacancy. The simplicity of this model allowed it to be a direct contribution in answering the research question, because it measured the relation between crime and house vacancy. This model also served as a reference point of comparison between other supporting variables such as population and housing market, that influence the connection between crime and house vacancy. These variables were iteratively added to the model where R^2 and adjusted R^2 were measured and compared. If a variable increased the performance of the model, it was added; however, if it decreased or had no impact on the model, a statistical significance test was performed to determine if it should be marked for future studies.

Ultimately, advanced models were compared against the baseline models to best predict the dependent variable. For a continuous target variable like HousVacant, Multiple Linear Regression was used to explore the relationship between the independent variables and the continuous target. The model’s performance was measured using Mean Squared Error (MSE) and R-squared (R²), which helped assess the accuracy of predictions and how well the model explained the variability in the target variable.

More advanced models, including Logistic Regression, Decision Trees (CART and C5.0), Random Forest, Naive Bayes, and Gradient Boosting, were also implemented to capture more complex patterns in the data. These models were particularly useful for predicting multi-class targets (High, Medium, Low). To make this possible, a Label Encoding library from Sklearn was used. This technique assigned each category a natural numerical order. The effectiveness of the models was assessed using metrics such as classification reports, ROC-AUC scores, and confusion matrices. Together, these regression and classification models provided a comprehensive understanding of the relationships between housing vacancy and its predictors, ultimately leading to the selection of the most accurate and reliable model.

Finally, the evaluation was further supported by visual tools such as Actual vs. Predicted (AvP) and error distribution graphs. These visualizations helped in evaluating the model’s accuracy and identifying potential biases. By combining these metrics and methods, the most suitable model for predicting housing vacancy was selected.

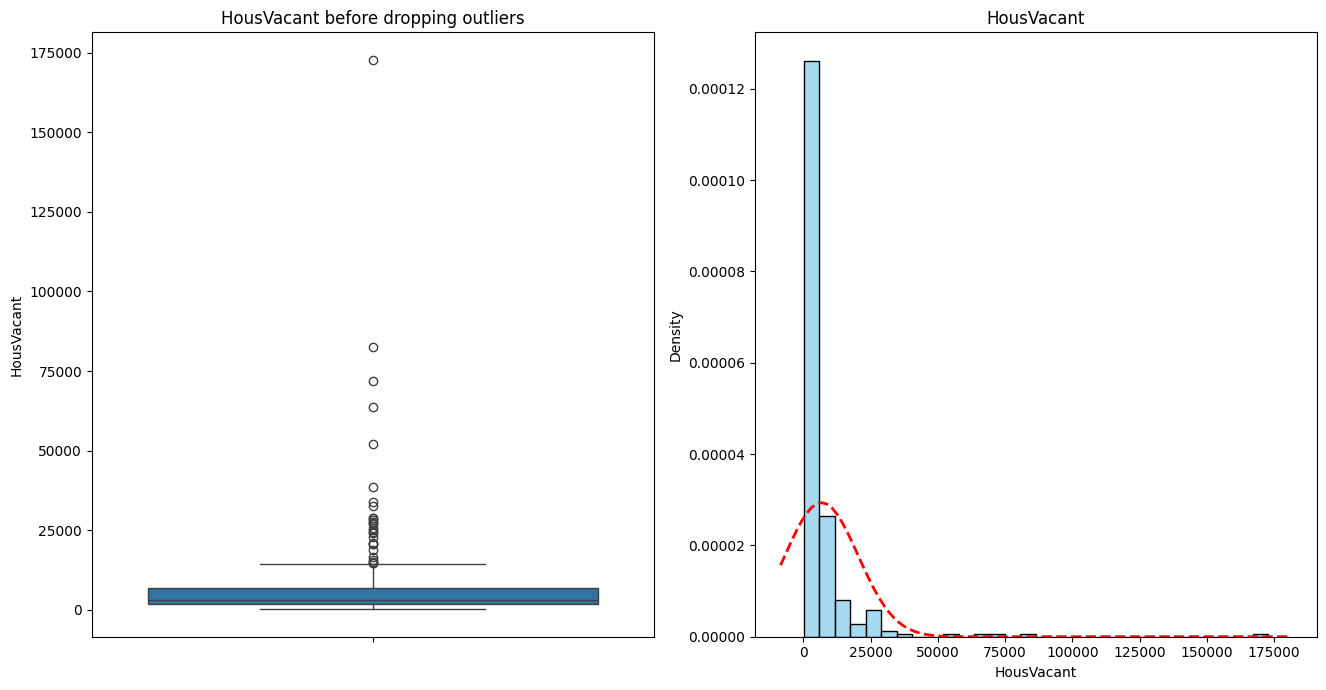
A cross-validation technique was applied to ensure the models were both reliable and generalizable. Specifically, 5-fold cross-validation was used to evaluate the performance of each classification model. This technique divides the dataset into five equal parts, or "folds." Each fold takes turns being used as the test set, while the remaining folds are used for training. The process is repeated for all five folds, and the results are averaged to provide a more accurate and consistent evaluation of the model's performance.

All the models were cross validated using this method. Their performance was assessed using key performance indicators appropriate for their type (Categorical/Continuous), offering a well-rounded understanding of how well each model performed across different subsets of the data. By using cross-validation, we ensured that the models were not overly reliant on a specific train-test split, reducing the chances of overfitting and providing a clearer picture of their overall predictive capabilities. These results helped in selecting the most effective model for predicting housing vacancy.

The models and metrics were computed by utilizing powerful python libraries such as SKLearn and Statistics. These tools provided more room for flexibility and precision. Instead of using a single type of model on a tool such as SAS, python provided the foundation for multiple models to run on a single loop, which trained the models, and computed all the necessary metrics to evaluate and validate the models. Additionally, data science tools such as visualization libraries are optimized in python, which assists in generating valuable graphs. This allows the research to better display and communicate the relationships between crime, vacancies, and the housing market.

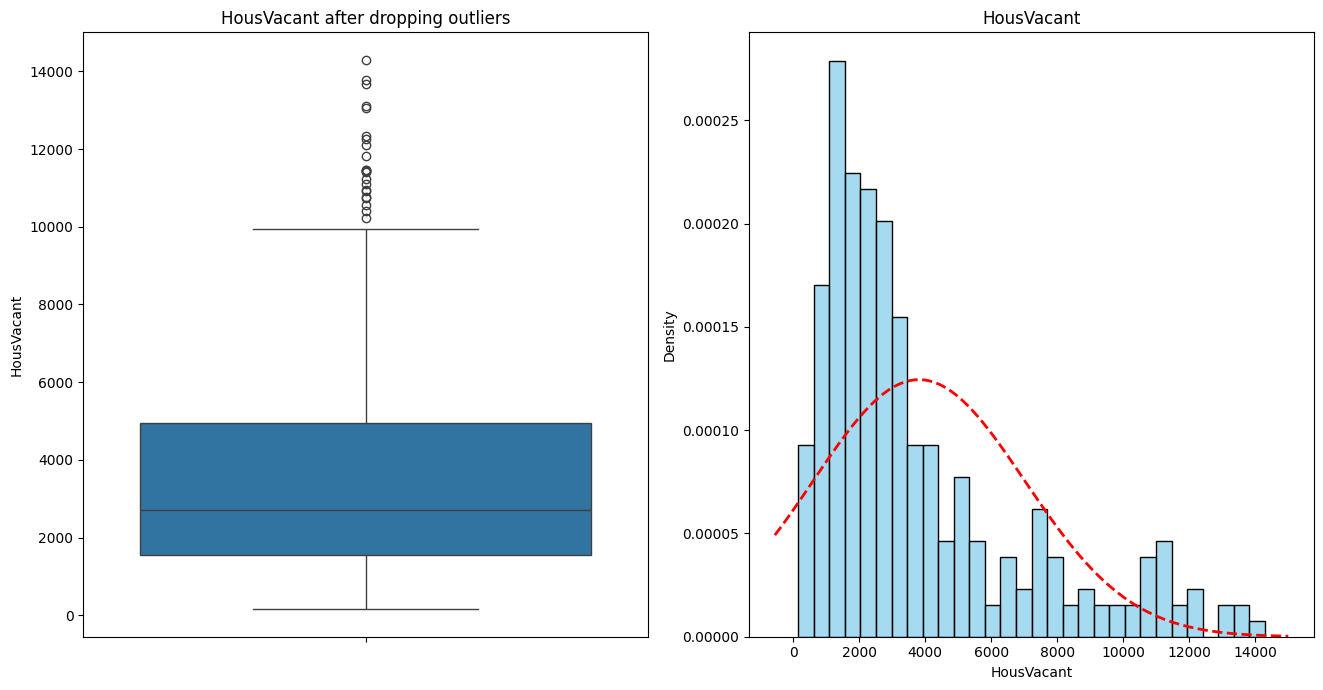
**Analysis**

As the dataset had many missing values, simply dropping these values would have resulted in 89% of the records to be deleted. To address this issue, all numeric variables have been extracted into a new data frame; All “?” values were converted to 0; and a KNN imputer technique was implemented. This technique identifies the nearest neighbors of a missing value and imputers their mean into the missing entry. This approach allowed us to keep all the observations while maintaining the dataset’s integrity.



*Figure [5] - Box plot and Distribution graph of HousVacant before removing outliers*

An example of how outliers were removed from the data set is reflected on Figure [5] which visualizes the 5 statistics of HousVacant. To remove the outliers, 2 additional statistical measures were calculated, upper bound and lower bound. All values above the upper bound were removed, and all the values below the lower bound were removed.



*Figure [6] - Box plot and Distribution graph of HousVacant after removing outliers*

In Figure [6] we can see that both the box plot and distribution are clearer after removing the outliers. However we still observe that HousVacant is naturally skewed to the right. The same was done on all other independent variables used in the model.

The variable ‘drug\_units\_ratio’ represents the ratio of officers assigned to drug units relative to the population density. This new feature provides a detailed understanding of law enforcement resource allocation relative to the population, which is key to analyzing its impact on the community and housing conditions. When the variable `drug\_units\_ratio` was introduced into the linear regression model, it resulted in a 3% increase in the adjusted R² score, indicating that it is better at explaining the variation in housing vacancies compared to using `NumKindsDrugsSeiz` (the number of different kinds of drugs seized). The strong connection between ‘drug\_units\_ratio’ and housing vacancies underscores its importance in predicting housing conditions.

*Table [2] - The distribution of the second (balanced) binning method for vanacy\_class*

| **Vacancy Class** | **Count** |
| --- | --- |
| High | 741 |
| Medium | 929 |
| Low | 545 |

Another variable that was created is vacancy\_class, which is a categorical representation of HousVacant. The first binning method, labeled all variables above the upper quartile as High, all variables below the lower quartile as Low, and everything in between as Medium. While this method accurately reflected the distribution of the data in the categories. It was greatly skewed towards the upper quartile. Hence, a second binning method was used; this method prioritized the balance between all 3 bins, with respect to the distribution of HousVacant. This method also ensured that the models are trained on balanced data. The distribution of the second method is reflected on Table [2].

The baseline model for the continuous variable, HousVacant, was created by predicting HousVacant against its mean value, 1748.4. The results of this baseline model are 1943.3 for mean absolute error, and 6502.4 for mean squared error. The large difference between mae and mse indicates the presence of outliers in the model. As mse penalizes the model for larger errors, mae will be used as a point of comparison between the continuous baseline model and advanced linear regression model.

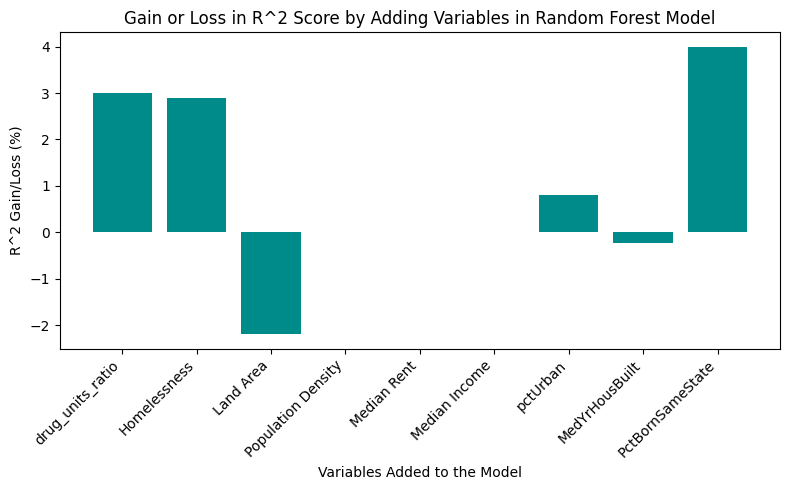
As for the categorical baseline model, HousVacant was first transformed into a categorical variable vacancy\_class. As shown in Table [2], this variable has three categories: Low, Medium, and High, with distributions of 545, 741, and 929, respectively. Since Medium is the mode, all vacancy\_class values were predicted as Medium in the baseline model. This model was evaluated, achieving an accuracy score of 41.95%. Advanced models will also have their accuracy scores computed and compared against this baseline.

As the research question investigates how non-violent crime predicts house vacancies, Larcenies, Burglaries, and Assigned drug units were the core of the model. To investigate how other variables influenced the relationship between crime and house vacancies, supporting variables such as population density, land area, homelessness, median rent, and median income were iteratively added to the model.

With the addition of the variable measuring homelessness, number of shelters, the adjusted R square of the model was **73.5%**. This **2.9%** increase indicates that the homelessness influences the relationship between house vacancies and crime.

The ensuing variables that were investigated are population density and land area. Both of these variables were added to the model that included homelessness. The addition of Land area decreased the performance of the model by **2.2%**, while population density had almost no impact on the performance of the model. To confirm the impact of population density on the model, a statistical T-test analysis was performed where the P-value was greater than 0.05, further proving that population density had no significant impact on the relationship between housing vacancies and crime. Following this conclusion, neither population density nor land area were added to the model

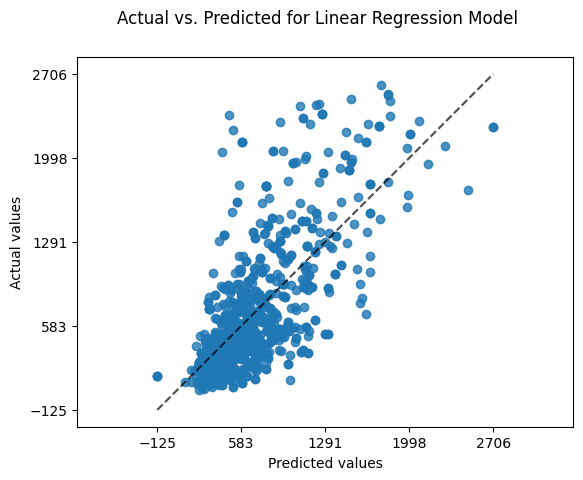
Lastly, the last two variables tested were median rent, and median income. The addition of median income and median rent individually, had almost zero impact on the model’s performance. To take a deeper look, another statistical T-test analysis was performed where both variables had an extremely low P-value indicating that there is a complex relationship between crime, house vacancies, homelessness, and the housing market.



*Figure [6] - R^2 gain and loss with the addition of each variable to the model.*

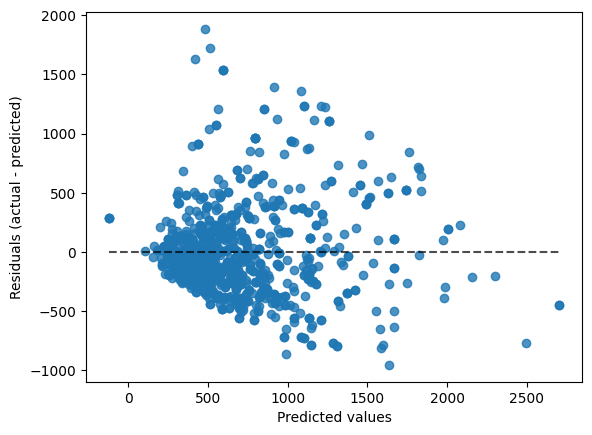
Figure [5] visually highlights the effectiveness of each variable on the model’s predictive performance by showing how much each addition influenced the R² score, either positively or negatively. This demonstrates the model's sensitivity to the relevance of specific features, reinforcing the importance of careful variable selection in predictive modeling.

The linear regression model provides a baseline for predicting HousVacant, by using factors like burglaries, larcenies, drug\_units\_ration, and NumInShelters, The model’s R^2 score of 48.4% means that nearly half of the variation in housing vacancies can be explained by these features. The RMSE of 398.50 suggests that, on average, the model’s predictions deviate by about 398 vacant houses from the actual values. Given that the mean value of HousVacant is 1748, this RMSE is 22.8% of the mean which is acceptable in the use case of this research.

****

*Figure [7] - Actual vs. Predicted scatter plot of Random Forest Model*

Figure [7] visualizes the actual values as the blue dots in the scatter plot, and predicted values as the dotted line. The actual values are more concentrated in the lower portion of the plot, suggesting that the model makes reasonably accurate predictions because these points are closer to the predicted line. However, as the value of HouseVacant increases the actual values are more spread out indicating that the number of accurate predictions decreases considerably.

****

**Root Mean Squared Error (MSE): 398.13738007940776**

**R-squared (R²): 0.483801135101147**

*Figure [7] - Residual plot of Random Forest Model*

The residuals plot further illustrates this point, revealing a noticeable spread and clear patterns, especially for higher predicted values of HousVacant. This indicates issues of heteroscedasticity, where the prediction errors become larger and more inconsistent as the predicted values increase. Overall the spread is nonuniform which suggests an inconsistent distribution of errors, highlighting the model’s limitations and impacting its use cases.

In comparison to the continuous baseline model, the linear regression model substantially outperforms the baseline. With an **MAE** of **398.14** and an **RMSE** of **398.14**, the linear regression model makes far more accurate predictions, significantly reducing average errors. Furthermore, its **R^2** score of **48.4%** demonstrates that the model explains nearly half of the variability in HouseVacant, a notable improvement over the baseline.

*Table [3] - K-fold average evaluation metrics of all categorical models*

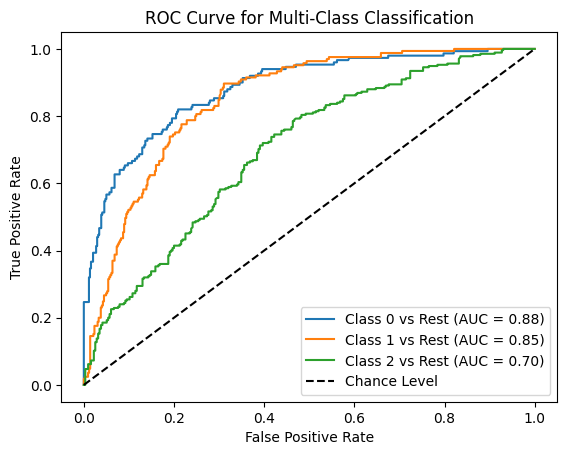
| **Model** | **Average Accuracy** | **Average Precision** | **Average Recall** | **Average F1-Score** | **Average ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | 66.3% | 68.9% | 66.3% | 65.9% | 82.5% |
| **Decision Tree** | 60.1% | 60.3% | 60.1% | 60.1% | 69.2% |
| **Random Forest** | ***69.3%*** | 69.9% | 69.3% | ***69.4%*** | 84.7% |
| **Naive Bayes** | 59.4% | 66.5% | 59.4% | 59.3% | 80.3% |
| **Gradient Boosting** | 68.8% | 69.3% | 68.8% | 68.8% | 84.8% |
| **XGBoost** | 66.7% | 67.3% | 67.2% | 68.6% | 83.8% |

The evaluation of classification models highlights clear differences in their performance. Random Forest and Gradient Boosting were the strongest models. Random Forest achieved the highest accuracy (69.3%) and F1-Score (69.4%), demonstrating its ability to handle complex patterns effectively while reducing overfitting by combining multiple decision trees. Gradient Boosting followed with an accuracy of 68.8% and an F1-Score of 68.8%, showing solid performance with potential for improvement through fine-tuning. Gradient Boosting also achieved the highest ROC-AUC (84.8%), reflecting its strong predictive capabilities.

Logistic Regression and XGBoost performed moderately well, with accuracies of 66.3% and 66.7%. However, their F1-Scores (65.9% for Logistic Regression and 68.6% for XGBoost) suggest challenges in balancing precision and recall. Logistic Regression served as a reliable starting point for predictions, while XGBoost showed potential for capturing more complex patterns.

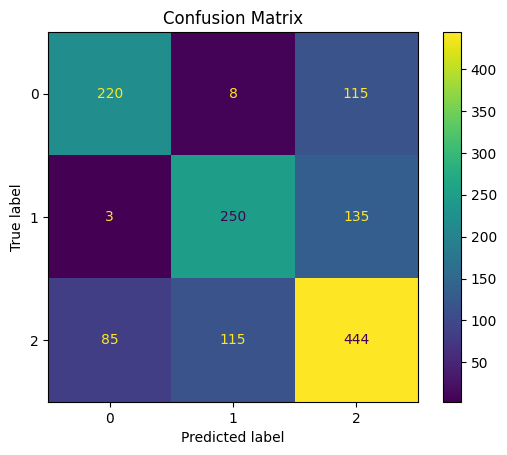
Decision Trees and Naive Bayes showed weaker performance, with Decision Tree achieving an accuracy of 60.1% and Naive Bayes having the lowest at 59.4%. Both models had difficulty handling the complexity of the dataset, which is shown by their lower F1-Scores (60.1% for Decision Tree and 59.3% for Naive Bayes). Naive Bayes struggled the most because its simple probabilistic approach could not handle the complexity of the data.

The use of 5-fold cross-validation ensured a thorough understanding by averaging results across different dataset splits, minimizing the effects of variability and bias. This analysis shows that Random forest and Gradient Boosting are the most reliable models for this task.



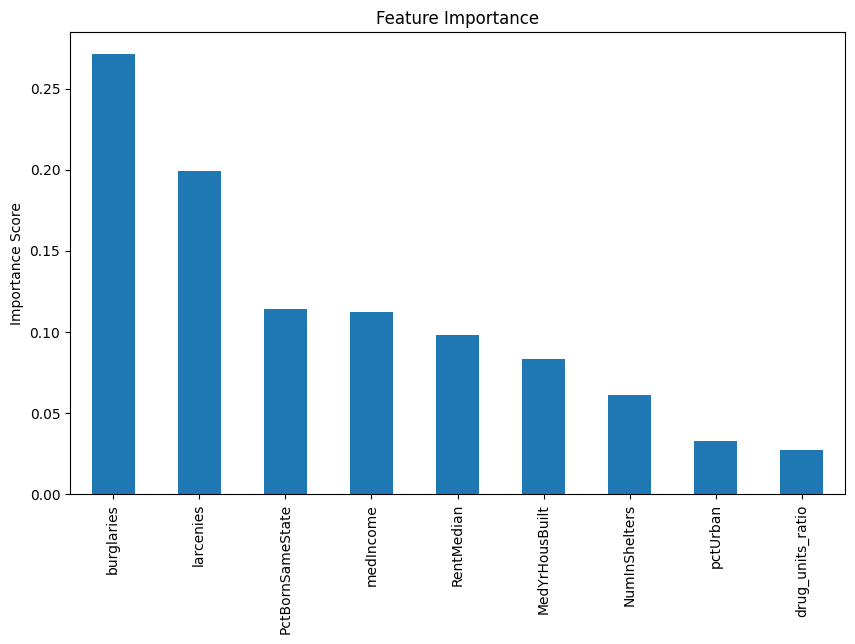
*Figure [8] - ROC Curve for all classes of HouseVacant in Random forest model*

The graph above shows ROC Curves for a multi-class classification problem, where each curve corresponds to one class. Class 0 represents areas with high housing vacancies (more than 900), Class 1 represents areas with low vacancies (under 300), and Class 2 represents medium vacancies (300-900). The curves evaluate each class using a “one-vs-rest” approach, plotting the true positive rate against the false positive rate. The AUC values: 0.88 for Class 0, 0.85 for Class 1, and 0.72 for Class 2 reflect the model’s performance. The high AUC values for Classes 0 and 1 indicate string performance, while Class 2 shows weaker predictive ability. The dashed line represents random guessing, highlighting the model’s overall effectiveness for most classes.

****

*Figure [9] - Confusion Matrix of Random Forest Model*

The confusion matrix shows how well the Random Forest model predicts housing vacancy levels: High(Class 0), Low(Class1), and Medium (Class 2). The numbers along the diagonal are the correct predictions: 220 for high, 246 for low, and 444 for medium. The off-diagonal numbers show errors, like 83 cases where High was predicted as Medium. Generally, the model had a high number of correct predictions, with classes 0, 1, 2 having 64%, 64%, 68% correct predictions respectively, which satisfies the expectations of this research.

****

*Figure [10] - Feature importance bar chart of all independent variables in the model*

This bar chart displays the feature importance scores generated by the Random Forest model, highlighting which features have the most significant influence on predicting housing vacancy levels. Burglaries and larcenies have the highest importance, meaning they play an important role in the model’s predictions. Other features like PctBornSameState, medIncome, and RentMedian are also important but have less influence. This chart helps identify which factors matter most for the model’s accuracy.

**Analysis Conclusion**

A key lesson learned from this analysis was the importance of carefully crafting features to enhance the performance of a model. For example, creating the ‘drug\_units\_ratio’ variable by merging existing features increased the model’s knowledge and accuracy. Another valuable lesson was the significance of early pre-processing, especially when handling outliers and addressing missing values. This step ensured the reliability of the results. Additionally, The advanced models proved to be useful, justifying its cost compared to the baseline models. Results from performing cross validation further increased our confidence in the models, and the claim of justified costs.

**Discussion**

This research builds on past studies by exploring the bidirectional nature of crime and residential vacancies through the combined lens of Broken Windows Theory and Social Disorganization Theory. These frameworks provided a foundation for understanding how visible neglect, such as vacant homes, signals weakened social cohesion, fostering conditions that perpetuate crime and further vacancies. Our study aimed to extend this understanding by analyzing these dynamics at the state level, providing a broader context for localized patterns observed in prior research.

However, several limitations emerged that shaped the scope of our analysis. The dataset, compiled from various sources over 29 years ago, posed challenges due to a significant number of missing values. Despite employing advanced imputation techniques, these gaps raise concerns about potential biases or oversights in the findings. Furthermore, although community-level data was available, it was too sparse to allow for detailed case studies. Aggregating the data to the state level mitigated this limitation, enabling a more comprehensive analysis while sacrificing some granularity. This trade-off highlights the need for more consistent and detailed data collection in future studies to better capture localized trends.

The ethical implications of this research were also carefully considered. Missing data introduced uncertainty, and we employed rigorous methodologies to minimize its impact. However, this research examines a complex societal relationship that is inherently influenced by systemic inequities and biases. For instance, data on crime and housing can inadvertently reflect or reinforce stereotypes. To mitigate these risks, we evaluated sensitive variables thoroughly to ensure that the analysis remained ethical, prioritizing accuracy and avoiding unintended consequences. This aligns with the theoretical principles guiding our research—highlighting the importance of fostering trust and equity in community interventions.

The findings of this research offer valuable insights for policymakers, urban planners, and community stakeholders. Policymakers, for instance, could use this information to develop targeted strategies for crime prevention that address underlying structural issues, such as investing in neighborhood revitalization programs to reduce visible neglect. Drawing from Broken Windows Theory, interventions could focus on repairing and repurposing vacant homes to send signals of stability and investment in affected neighborhoods. Similarly, insights from Social Disorganization Theory emphasize the importance of fostering social cohesion through community-based programs and improved housing policies to disrupt the cycle of crime and abandonment.

Finally, future research could build on these findings by utilizing more recent, community-level data to explore the nuances of crime and housing vacancies in greater detail. Expanding this work could provide actionable insights into how state-level policies, such as tenant protections and housing subsidies, influence these relationships across various socioeconomic and demographic contexts. By integrating more contemporary datasets and accounting for localized factors, future studies could improve the robustness and relevance of these analyses, ultimately paving the way for more effective, equitable solutions.

**Conclusion**

This research successfully achieved its objective of examining the relationship between crime, residential vacancies, and housing market conditions at the state level, while building on the foundational frameworks of Broken Windows Theory and Social Disorganization Theory. By situating this analysis within these theoretical contexts, the study illuminated how structural and social factors contribute to cycles of neglect, crime, and abandonment. Specifically, it identified non-violent crimes, homelessness, and cultural factors as significant drivers of residential vacancies, underscoring the role of societal dynamics over purely economic influences. In contrast, housing market factors, such as median home prices, were found to have a less pronounced impact on residential vacancies, suggesting that social stability plays a more critical role in these patterns than previously assumed.

While the study made valuable contributions, it also faced several challenges. The dataset, being over 29 years old and featuring significant missing values, limited the analysis in terms of both temporal relevance and granularity. Despite employing advanced imputation techniques and aggregating data to the state level, these constraints highlight the need for more up-to-date and community-specific data to fully capture the nuances of crime and housing dynamics. Nevertheless, the analysis provided a comprehensive exploration of these interconnected issues, contributing new insights to an ongoing discourse on urban decay and renewal.

Ultimately, this research reinforces the importance of addressing the social and structural underpinnings of residential vacancies and crime. The findings suggest that policymakers, urban planners, and community leaders should prioritize strategies that enhance social cohesion, address homelessness, and mitigate the effects of non-violent crimes. These insights not only contribute to the academic understanding of the bidirectional relationship between crime and housing but also offer practical pathways for fostering more stable, resilient communities. Future research that incorporates modern, community-level data and explores the role of emerging factors, such as gentrification and economic inequality, could further advance this field and support more effective, evidence-based policymaking.

**References**

14th United Nations Congress on Crime Prevention and Criminal Justice (2020). *Comprehensive strategies for crime prevention towards social and economic development.* United Nations. <https://unis.unvienna.org/pdf/2021/Crime_Congress/04_integrated_approaches_criminal_justice_FINAL.pdf>

Boessen, A., & Chamberlain, A. W. (2017). *Neighborhood crime, the housing crisis, and geographic space: Disentangling the consequences of foreclosure and vacancy. In Journal of Urban Affairs 39*(8). Informa UK Limited. <https://doi.org/10.1080/07352166.2017.1310558>

Boggess, L. N., & Hipp, J. R. (2010). *Violent crime, residential instability and mobility: Does the relationship differ in minority neighborhoods? - journal of quantitative criminology.* SpringerLink. <https://link.springer.com/article/10.1007/s10940-010-9093-7#citeas>

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2012). *The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis*. Justice Quarterly, *31*(4). <https://doi.org/10.1080/07418825.2012.673632>

Brunson, R. K., & Weitzer, R. (2008). Police relations with black and white youth in different urban neighborhoods. *Urban Affairs Review, 43*(6). <https://doi.org/10.1177/1078087408323792>

Bureau of Justice Statistics. (2022). *Justice expenditure and employment extracts, 2021*. U.S. Department of Justice. Retrieved from <https://bjs.ojp.gov>

Chaskin, R. J., & Joseph, M. L. (2014). *Integrating the inner city: The promise and perils of mixed-income public housing transformation.* University of Chicago Press.

Chen, X., & Rafail, P. (2020). Do Housing Vacancies Induce More Crime? A Spatiotemporal Regression Analysis. *Crime & Delinquency, 66*(11). <https://doi.org/10.1177/0011128719854347>

Cozens, P., Hillier, D. and Prescott, G. (2001), Crime and the design of residential property – exploring the perceptions of planning professionals, burglars and other users: Part 2, *Property Management, 19*(4). <https://doi.org/10.1108/EUM0000000005784>

Cui, L., & Walsh, R. (2014). *Foreclosure, vacancy and crime*. NBER. <https://www.nber.org/papers/w20593>

Cullen, J. B., & Levitt, S. D. (1999). *Crime, Urban Flight, and the Consequences for Cities. In Review of Economics and Statistics 81*(2). MIT Press - Journals. <https://doi.org/10.1162/003465399558030>

Desmond, M. (2016). *Evicted: Poverty and profit in the American city*. Crown Publishers/Random House. <https://psycnet.apa.org/record/2016-12857-000>

Dugan, L. (1999). *The Effect of Criminal Victimization on a Household’s Moving Decision.* Criminology, 37(4). <https://doi.org/10.1111/j.1745-9125.1999.tb00509.x>

Farrall, S., Hunter, B., Sharpe, G., & Calverley, A. (2014). *Criminal careers in transition: The social context of desistance from crime.* Hull Repository. <https://hull-repository.worktribe.com/output/1934822/criminal-careers-in-transition-the-social-context-of-desistance-from-crime>

Farrall, S. D., Jackson, J., & Gray, E. (2009). *Social order and the fear of crime in contemporary times*. Oxford University Press. <https://books.google.com/books?hl=en&lr=&id=nesSwIIJoagC&oi=fnd&pg=PR13&dq=Social+Order+and+the+Fear+of+Crime+in+Contemporary+Times&ots=CeVmcQENCw&sig=i6uRablyvjZ4uqSIQNdpC4mBzNc#v=onepage&q=Social%20Order%20and%20the%20Fear%20of%20Crime%20in%20Contemporary%20Times&f=false>

FBI. (2024). Crime data explorer: Crime trends. U.S. Department of Justice.<https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/crime-trend>

Focareta, D. (2024). *Home vacancy statistics 2024.* ConsumerAffairs. [https://www.consumeraffairs.com/finance/home-vacancy-statistics.html‌](https://www.consumeraffairs.com/finance/home-vacancy-statistics.html%E2%80%8C)

Michigan News (2020). Detroit housing shortage, evictions set stage for covid-19 housing crisis. Gerald R. Ford School of Public Policy. <https://fordschool.umich.edu/news/2020/detroit-housing-shortage-evictions-set-stage-covid-19-housing-crisis>

Morenoff, J. D., Sampson, R. J., & Raudenbush, S. W. (2001). *Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence*. Criminology, 39(3). <https://doi.org/10.1111/j.1745-9125.2001.tb00932.x>

Olajide, S., & Lizam, M. (2017). *Determining the impact of residential neighbourhood crime on housing investment using logistic regression.* SSRN. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2949339>

Porter, L. C., De Biasi, A., Mitchell, S., Curtis, A., & Jefferis, E. (2018). *Understanding the Criminogenic Properties of Vacant Housing: A Mixed Methods Approach*. Journal of Research in Crime and Delinquency, 56(3). <https://doi.org/10.1177/0022427818807965>

Rainiero, N. (n.d.). Neighborhood Revitalization Tax Credit Program. Housing & Community Development Network of New Jersey. <https://www.hcdnnj.org/neighborhood-revitalization-tax-credit>

Rizzo, M. J. (1979). *The Effect of Crime on Residential Rents and Property Values.* The American Economist, 23(1). SAGE Publications. <https://doi.org/10.1177/056943457902300103>

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing Disorder: Neighborhood Stigma and the Social Construction of “Broken Windows”. *Social Psychology Quarterly, 67*(4). <https://doi.org/10.1177/019027250406700401>

Shaw, C. R., & McKay, H. D. (1942). *Juvenile delinquency and urban areas.* The University of Chicago Press. <https://psycnet.apa.org/record/1943-00271-000>

Skogan, W. G. (2006). *The promise of community policing in America: A response to the challenges of crime and disorder*. *Crime and justice: A review of research*. University of Chicago Press. <https://www.skogan.org/files/An_Overview_of_Community_Policing.pdf>

Spelman, W. (2002). Abandoned buildings: Magnets for crime? *Journal of Criminal Justice, 30*(5). <https://doi.org/10.1016/S0047-2352(02)00149-X>

Van Doren, P. (2016). *Public housing and crime*. Regulation, 39(2). Gale Academic OneFile, [link.gale.com/apps/doc/A457972383/AONE?u=char69915&sid=googleScholar&xid=af960d7e](http://link.gale.com/apps/doc/A457972383/AONE?u=char69915&sid=googleScholar&xid=af960d7e). Accessed 3 Dec. 2024.

Whitaker, S., & Fitzpatrick, T. J. (2012). *The impact of vacant, tax-delinquent, and foreclosed property on sales prices of neighboring homes*. Federal Reserve Bank of Cleveland. Retrieved from <https://www.clevelandfed.org>

Wilson, J. Q., Kelling, G.L. (1982). Broken windows. *Critical issues in policing: Contemporary readings.* <http://www.theatlantic.com/politics/crime/windows.htm>