Impact of Crime and Property Quality on Rental Prices and Tenant Satisfaction in Los Angeles

Team Member: Fan-Chen Fu / fanchenf@usc.edu / 5980024871

1 Introduction

1.1 Background

In the summer of 2024, a young USC student was attacked in a "luxury" apartment in downtown Los Angeles – the same apartment I had been living in for a year. This incident deeply unsettled me, as the apartment was marketed as a secure residence with 24-hour security and gated access. It raised a critical question about the relationship between safety measures, property quality, and rental costs, prompting me to wonder whether the high rent reflected the value and security provided. Motivated by these concerns, this project examined how crime count, property ratings, and amenities impacted rental prices and tenant satisfaction in Los Angeles.

1.2 Research Question

While the frequency of crime has been known to impact property values, the extent to which it influenced rental prices and tenant satisfaction remained less explored, especially in a city as diverse and expansive as Los Angeles. Additionally, the role of safety-related amenities in mitigating the effects of neighborhood crime and their perceived value to tenants warranted further investigation. To address these complexities, this project sought to answer the following research questions: **(1) How do neighborhood crime rates influence rental prices in Los Angeles? (2) Do safety-related amenities in apartment complexes correlate with crime rates in the surrounding area? (3) How do property ratings reflect tenant satisfaction with neighborhood crime and nearby amenities?** The project uncovered key factors shaping rental market dynamics by addressing these questions.

2 Datasets

Focusing on dynamic relationships between rental price, satisfaction, and crime count, the project utilized three key datasets: The **Apartments.com (Apt-Raw)** dataset was scraped from Apartments.com; the **City of Los Angeles Crime Data (LA-Crime-Raw)** dataset was downloaded from the official resources. Lastly, **the United States Zip Codes dataset (US-Zip)** provided was used to match the rental and crime data by postal codes and city names. More details are shown below.

2.1 Apt-Raw

The property data was sourced from Apartments.com, one of the nation's most comprehensive online rental marketplaces with extensive rental listings. To ensure thorough data collection for Los Angeles County, the official definition from the County of Los Angeles was applied to identify the 88 incorporated cities. These city names were used as keywords to search on Apartments.com, and data for all listed properties in those cities was directly scraped with Python code. I referred to this dataset as "Apt-Raw" in this report. It contains 13,246 rows with columns including "property name," "address (with postal code)," "price," "property link," "amenities," "property rating," and "review count." Notably, property ratings provide valuable insight into tenant satisfaction and can reflect property quality to some extent, making them helpful in analyzing the correlation between satisfaction, price, and safety factors.

2.2 LA-Crime-Raw

The City of Los Angeles Crime Data provides this publicly available dataset in CSV format, referred to as"LA-Crime-Raw" in this report, with records accumulated from 2020 onward to **November 9, 2024** (the download date). It contains 217.774 trillion rows transcribed from original paper crime reports. The dataset includes a wide range of columns such as "record number," "date of occurrence," "crime committed," "location address," "latitude," and "longitude." This extensive dataset allows for an in-depth analysis of crime patterns, crime count, and their relationship with rental properties.

2.3 US-Zip

The dataset was downloaded from UnitedStatesZipCodes.org in CSV format, including all postal (zip) codes and primary cities used by USPS. This dataset was referred to as "US-Zip" in the report, and it contained 2,108,897,698 rows, along with associated information such as "primary timezone," "area code," and "counties." The dataset was used as the reference to match the Apt-Raw with the LA-Crime-Raw on city name and postal code. This matching ensures consistent and accurate integration of property and crime datasets.

3 Data Cleaning and Pre-processing

3.1 Filter US-Zip to Keep Only Los Angeles Zip Codes (LA-Zip)

Since the US-Zip datasets contain data around the States, I wanted to keep only the zip codes of Los Angeles County, ensuring accurate geographical mapping for subsequent analyses. A predefined list of 88 incorporated cities within Los Angeles County was used as a filter based on the official information from County of Los Angeles. To begin with, I identify ZIP codes in California (state == 'CA'). Secondly, I narrowed the data to include only records whose primary city matches one of the listed Los Angeles County cities. The filtered dataset provides precise associations between cities and their respective ZIP codes, ensuring compatibility with other datasets, such as crime and rental property data. The cleaned and filtered data was then saved to a CSV file and referred to as "LA-Zip" in this report, forming a reliable baseline for integrating location data in the project.

3.2 Standardize LA-Crime-Raw and Add Detailed Prices, Locations, and Amenities Columns (Clean-Rental)

This phase involved standardizing, extracting, and enriching the datasets to ensure consistency and usability for analysis.

Minimum Prices: Initially, I handled inconsistencies in the price column, converted non-numeric entries like "Call for Rent" to NaN, and split price ranges into minimum and maximum values. Considering Apartments.com's reported average rent in Los Angeles is $2,155 per month for 1B, and the typical variance in rent ranges across floorplans (e.g., studio, 1B, 2B, and 3B), I applied a $7,000 threshold to filter out properties with unreasonably high **minimum prices**, assuming even a 3B property rented by a single tenant would fall below this limit.

Postal codes: I extracted postal codes from addresses to enable matching with the LA-Zip. First, to achieve this, I parse the postal code from the address string by isolating the last five digits of each address. After cleaning and verifying the postal codes, I performed a merge operation between the dataset and the LA-Zip, using the postal code as the key, allowing me to append a new "city_name" column, representing the primary city associated with each zip code. Finally, rows without valid postal codes or city matches were excluded to ensure the integrity and accuracy of the location mapping. This step was critical for associating rental properties with their respective cities, enabling further analysis of rental prices and crime data at a granular, city-specific level.

Security-Related Amenities: I identified and categorized security-related features using a set of predefined keywords, such as "security," "gated," "patrol," and "controlled access," creating columns to count the occurrence of each keyword and calculated the total security-related amenities per property. Additionally, both columns were standardized for property ratings and review counts by converting non-numeric entries to zeros.

Then, I validated URLs, removed duplicates, and enriched the dataset with features like total security amenities and accurate location data. The cleaned dataset was saved for further analysis and referred to as "Clean-Rental" in this report, providing a solid foundation for meaningful exploring.

3.3 Obtain Postal Code Information for LA-Crime-Raw and Categorize Cime Category (Clean-Crime)

First, I filtered LA-Crime-Raw to include records from the specified six-month period (May 9, 2024, to November 9, 2024), which was efficient enough to get all trends and patterns needed, and removed rows with missing or invalid latitude or longitude values and duplicate entries to ensure data consistency. To enrich the dataset with postal codes, I utilize the geopy library, applying reverse geocoding to convert latitude and longitude coordinates into postal codes. A randomized user agent and delay were implemented to comply with API rate limits and avoid server restrictions during the geocoding process. Postal codes were appended to the dataset, and rows without valid postal codes were excluded. The enriched clean dataset was then saved to a new CSV file with accurate location data for the exact period. Next, the dataset was merged with the filtered Los Angeles zip codes dataset to append a "city_name" column, enabling accurate geographical context for each crime record.

Secondly, the crime descriptions were standardized by converting them to lowercase and stripping whitespace. A mapping dictionary was applied to categorize crimes into broader categories, such as "theft," "assault," "sexual offenses," and "others." This simplification gave more precise insights into crime patterns by grouping similar offenses under unified categories.

Finally, the cleaned dataset with all needed metrics was saved as a new CSV file and referred to as "Clean-Crime" in this report. It is ready for further analysis of crime trends and their relationships with other variables.

4 Data Analysis and Visualization

The following hypotheses were formulated based on the research question: **(1) H1**: Higher neighborhood crime counts are associated with lower rental prices. **(2) H2**: Properties with more security-related amenities have higher rental prices, regardless of surrounding crime rates. **(3) H3**: High-crime neighborhoods have a higher prevalence of security amenities than low-crime areas. **(4) H4**: Tenant satisfaction is positively correlated with the presence of security amenities, especially in high-crime neighborhoods.

4.1 Crime and Price Correlation Analysis

I conducted a detailed analysis using linear and advanced regression models to explore the relationship between crime counts and rental prices across Los Angeles neighborhoods, with data grouped at the city and postal code levels.

4.1.1 Initial Analysis: Crime and Rental Price Correlation by City

I first grouped data by **city** and calculated the average minimum rental price and total crime count. However, the extremely high crime count in Los Angeles City (28,242 records), far exceeding that of other cities, skewed the linear regression results (**Fig1**). Comparing the results in Fig1 and Fig2, after removing Los Angeles City, we can observe a better correlation between crime count and the average minimum rental price. The scatter plot showed **a negative relationship between crime count and average minimum rental price, supporting H1.** (**Fig2**)

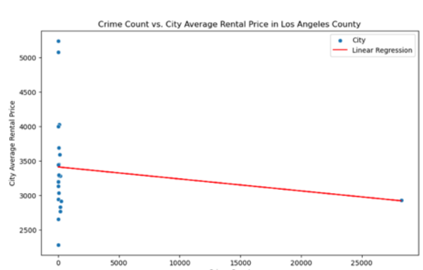


Fig 1. Crime Count and Average Minimum Rental Price by City in Los Angeles County

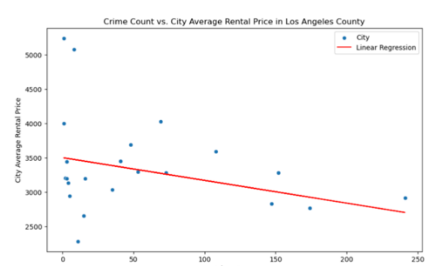
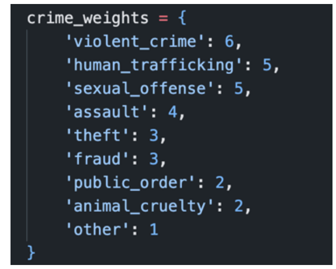


Fig 2. Linear Regression Model's Result Without Los Angeles City

4.1.2 Focusing on Los Angeles City: Postal Code Grouping

Recognizing the importance of Los Angeles City for this analysis, I shifted focus to its neighborhoods, grouping data by postal codes. After removing outliers with IQR in rental prices, linear regression demonstrates a correlation between the features, while the R^2 score is only **0.2011,** reflecting the possibility of a multiple features correlation to the rental price. Therefore, I will construct relevant experiments in the following section. (**Fig 3**)

Fig 3. Crime Count and Average Minimum Rental Price by Postal Code in Los Angeles City



```
crime_weights = {
    'violent_crime': 6,
    'human_trafficking': 5,
    'sexual_offense': 5,
    'assault': 4,
    'theft': 3,
    'fraud': 3,
    'public_order': 2,
    'animal_cruelty': 2,
    'other': 1
}
```

Fig 4. Severity Scores for Each Crime Category

4.1.3 Feature Enrichment and Random Forest Modeling

To enhance the linear regression model's performance, I incorporated additional features, including "property ratings," "review counts," and "total security amenities" from the Clean-Rental dataset and appended multiple crime categories from the Clean-Crime dataset. Additionally, to reflect the severity of different crime categories rather than treating all crimes equally, I introduced a "weighted crime score," which assigns higher values to more severe offenses based on the FBI Crime Data Explorer definition. (**Fig 4**).

The feature enrichment significantly performed a R^2 score of **0.7281** from the random forest regression model, indicating that the enriched feature set effectively explains the variance in rental prices. In Addition, the feature importance analysis from random forest regression echoed the result by highlighting that the **average total security amenities** emerged as the most influential feature (**Fig 5**); this supports **H2** by emphasizing the strong impact of safety-related property characteristics in driving rental value. Interestingly, the **weighted crime score** ranked higher than several specific crime categories, validating the utility of this composite metric in housing market analysis.

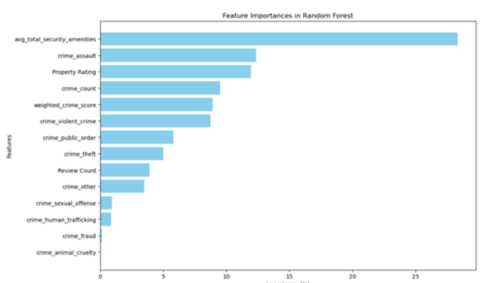


Fig 5. Feature Importance of the Random Forest Model

	Feature	Importance	Importance (%)
3	avg_total_security_amenities	0.283393	28.339287
6	crime_assault	0.123664	12.366386
1	Property Rating	0.119533	11.953317
0	crime_count	0.095219	9.521908
4	weighted_crime_score	0.088954	8.895450
13	crime_violent_crime	0.087480	8.748041
10	crime_public_order	0.058158	5.815759
12	crime_theft	0.049973	4.997277
2	Review Count	0.039096	3.909633
9	crime_other	0.035110	3.511010
11	crime_sexual_offense	0.009410	0.941037
8	crime_human_trafficking	0.008708	0.870822
7	crime_fraud	0.001301	0.130074
5	crime_animal_cruelty	0.000000	0.000000

Fig 6. Importance Score for Each Feature

4.1.4 Advanced Modeling: Optimizing for Accuracy

I applied feature selection based on importance scores from the random forest model, retaining only features with an importance score above 5 %. (**Fig 6**) This step reduced noise and improved model efficiency by focusing on the most impactful predictors. According to **Table 1**, with the updated feature, the random forest regression improved the R^2 score from 0.7281 to **0.7491,** proving the selection is effective. Furthermore, I also implemented other regression models for comparison, including XGBoost and CatBoost. Among the models, **XGBoost** demonstrated the best performance, achieving a R^2 score of **0.7531** and a low RMSE. It demonstrated its strength in handling larger errors and explaining variance, making it the most balanced and robust model to predict rental prices in the analysis. Its superior accuracy showed its ability to capture detailed interactions between features, such as how security amenities and crime patterns jointly influence rental prices.

Model	RF	XGBoost	CatBoost
MAE	**261.7168**	268.5743	314.982
RMSE	342.1151	**339.3696**	384.6368
MAPE	**10.4865**	10.6718	12.9763
R-squared	0.7491	**0.7531**	0.6828

Table 1. Result and Comparison of Multiple Regression Model

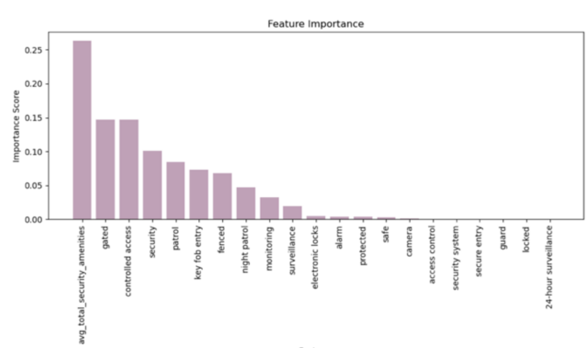


Fig 7. Feature Importance of the Random Forest Model in High-Crime Area

4.2 Amenity and Crime Pattern Analysis

4.2.1 Amenity Prevalence and Crime Levels

To determine whether an area is a high-crime neighborhood, I used the median of the weighted crime score across all postal codes as a threshold, and postal codes with higher weighted crime scores were labeled as "high-crime" areas, otherwise were labeled as "low-crime" areas.

With the precise definition, to analyze the relationship between safety-focused amenities and high-crime neighborhoods, firstly, I applied logistic regression using the average total security amenities as the sole feature, achieving an accuracy of **76%** with a positive coefficient of 1.206, indicating that an increase in security amenities slightly raises the likelihood of an area being classified as high-crime.

What's more, leveraging multiple features, including the average total security amenities and all security-related amenity columns, the random forest model achieved a significantly high accuracy of **86%**. Moreover, Random Forest performs well for both classes, particularly for high-crime areas, with a recall of 0.92, indicating that it successfully identifies most high-crime areas. Also, the random forest model revealed feature importance in that the total aggregate count of security-related amenities was the most important predictor, emphasizing that overall amenity presence plays a stronger role than individual features. (**Fig 7**) Physical barriers like gated access and controlled entry ranked second and third in importance, underscoring their prioritization in high-crime neighborhoods.

4.2.2 Distribution of Amenities Across Crime Levels

Fig 8 illustrates the average total amenities count across different crime levels, highlighting that high-crime neighborhoods tend to feature more security amenities on average. This discrepancy suggests a direct correlation between higher crime levels and the prevalence of security-focused amenities, reflecting landlords' efforts to address safety concerns and attract tenants in high-crime areas, which also directly supported **H3**.

On top of that, the pie chart in **Fig 9** displayed the aggregated count of each amenity in the high-crime areas, showing that these areas commonly feature a higher prevalence of physical safety measures such as gated access and security systems. The result aligned closely with the feature importance results from the random forest model, reinforcing the relationship between crime levels and security infrastructure.

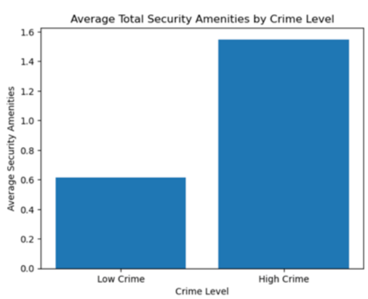


Fig 8. Average Total Amenities Count Across Crime Level

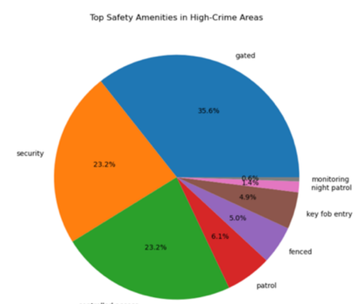


Fig 9. Safety Amenities in High-Crime Areas

4.2.3 Specific Crime Types and Amenity Needs

To answer the question, "What specific crime types (e.g., theft, assault) drive the need for certain amenities?" I constructed a heatmap graph capturing the average presence of each amenity across crime categories. (**Fig 10**) The ranked results and a heatmap visualization revealed that severe crimes, such as sexual offenses and human trafficking, significantly increase the presence of specific safety measures. For these crime categories, gated access, controlled entry, and security were prioritized, indicating their importance in mitigating severe criminal activity.

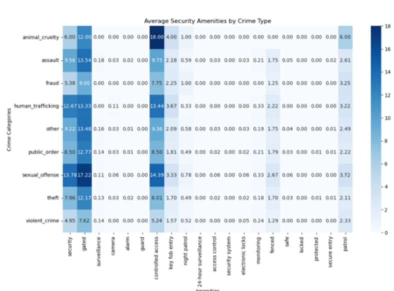


Fig 10. Average Security Amenities by Crime Type

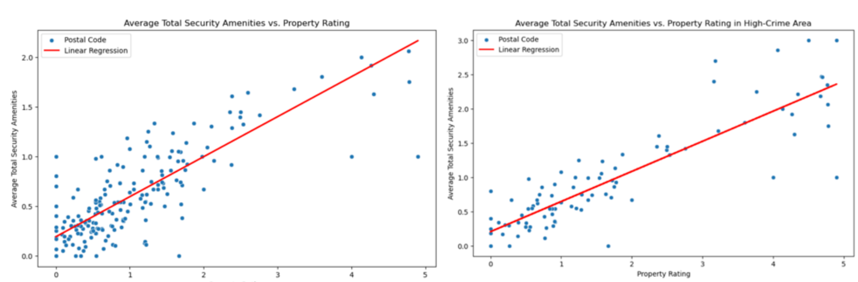


Fig 11. Average Total Security Amenities and Property Rating (All vs High-Crime Areas)

4.3 Safety Score and Satisfaction Prediction

4.3.1 Tenant Satisfaction and Security Amenities

I applied "property rating" as the definition of the tenant satisfaction value. The relationship between tenant satisfaction and security amenities was analyzed using linear regression models. For all neighborhoods grouped by postal codes, a linear regression examining average total security amenities and property ratings yielded a R^2 score of 0.5707, stating a moderate positive correlation. However, the correlation became more robust when focusing on high-crime areas, with a R^2 score of **0.8857** (**Fig 11**). This result examined **H4** to be positive and underscores that security amenities are particularly valued in high-crime areas, directly influencing tenant satisfaction.

4.3.2 Feature Importance in High-Crime Areas

Focusing on high-crime areas, a random forest regression model achieved a R^2 score of **0.9081**, emphasizing the unparalleled predictive power of total security amenities for tenant satisfaction. When focusing on each security-related amenity as a feature, the model scored a R^2 score of **0.6945,** and feature importance analysis revealed that controlled access, gated amenities, and key fob entry were the most critical amenities, providing visible and tangible security measures that tenants value. (**Fig 12**) Supporting features, such as "fenced" and "night patrols," also played a role, while "camera" and "alarm" were deemed less impactful, likely due to their reactive rather than preventive nature.

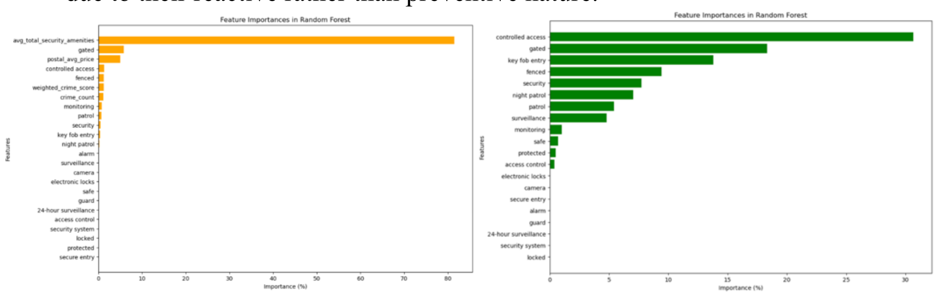


Fig 12. Amenities Feature Importance of the Random Forest Model in High-Crime Area

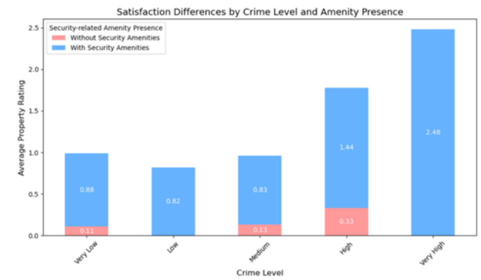


Fig 13. Satisfaction Level by Crime Level and Amenity Presence

4.3.3 Difference in Satisfaction Across Crime Levels and Amenity Presence

Satisfaction differences across crime levels were explored by categorizing neighborhoods into five logarithmic bins: **Very Low, Low, Medium, High, and Very High.** The stacked bar in **Fig 13** showed that properties with security amenities consistently achieved higher ratings across all crime levels than those without; specifically, in Very High crime areas, properties with amenities recorded the highest average property rating of **2.48**, compared to significantly lower ratings for properties without amenities. In safer neighborhoods, the difference was less pronounced, suggesting that tenants place greater importance on security features when the perceived risk is higher.

4.4 Conclusion

Analyzing crime, property quality, and tenant satisfaction in Los Angeles's rental market reveals several key insights highlighting the interplay between safety, amenities, and rental prices. To begin with, crime counts negatively impact rental prices, with higher crime levels correlating to reduced property value, although the relationship is influenced by additional factors such as neighborhood characteristics. Secondly, the result revealed that security-related amenities emerged as a critical factor, both in enhancing rental prices and improving tenant satisfaction, particularly in high-crime neighborhoods where tenants place a premium on visible safety features like gated access and controlled entry.

In Addition, the analysis offers valuable insights from various stakeholders. Tenants can prioritize properties with robust security features to balance affordability and safety, especially in high-crime areas. For landlords, investing in safety amenities enhances tenant retention and justifies higher rents, while targeted upgrades and strategic marketing boost property appeal. Real estate developers and analysts can use these patterns to design relevant properties and forecast rental trends. Overall, the findings provide a framework for improving living standards, maximizing property value, and guiding data-driven decisions in the housing market.

5 Changes from Original Proposal

In the original proposal, the goal was to analyze neighborhoods across the entire Los Angeles County. However, during the data analysis phase, it became evident that the crime count in Los Angeles City was significantly higher than in other areas, with 28,242 records within six months compared to an average of just 55 in other cities. This stark disparity created challenges in accurately analyzing the relationship between crime counts and the average minimum rental price across cities in Los Angeles County. Given the prominence of Los Angeles city as one of the most dangerous areas in the United States, the analysis shifted to focus primarily on neighborhoods within Los Angeles city. These neighborhoods were grouped based on postal codes to ensure a more meaningful and focused examination.

6 Future work

Expanding the dataset to include a broader time frame or additional cities could provide a more comprehensive analysis of long-term trends and regional variations. Incorporating socioeconomic factors, such as income levels, employment rates, and education, would offer more profound insights into how external conditions influence rental markets. Advanced modeling techniques, such as neural networks or ensemble models, could be explored to enhance predictive accuracy further. Additionally, sentiment analysis of tenant reviews by fine-tuned BERT model on Huggingface to provide qualitative insights into satisfaction beyond numerical ratings. These extensions would enhance the robustness and applicability of the findings, making them even more valuable for decision-making in the housing market.