Correlation of Tucson Crime Rate with Streetlights and Household Income

MohammadHossein Rezaei
University of Arizona
mhrezaei@arizona.edu

Stephin Tomson
University of Arizona
stephintomson@arizona.edu

Calvin Briscoe
University of Arizona
calbriscoe@arizona.edu

Connor Kippes
University of Arizona
connorkippes@arizona.edu

Abstract

Crime is a major concern in urban areas and has significant implications for public safety and quality of life. The factors contributing to crime rates are complex and multifaceted and discovering them is essential. In this project, we focus on Tucson, Arizona, and use publicly available datasets to run a data-driven analysis of crime rates. Specifically, we investigate the correlation between streetlight density, neighborhood median income, and property-related crime rates. We build regression and classification models to predict crime rates and identify high-risk areas. Our results indicate that there is a significant correlation between household income and crime rates, with lower-income neighborhoods experiencing higher crime rates. We also find that streetlight density is a predictor of crime rates, but the correlation is weaker than that of income. Feature importance analysis also confirms that both neighborhood income and and parcel density are strong predictors of property crime risk. This project is submitted as a final project for the course *CSc 380: Principles of Data Science* at the University of Arizona.

1 Introduction

Crime is a critical issue in urban areas, which affects the quality of life for residents and the overall safety of communities. Crime is a complex phenomenon influenced by various factors, including socio-economic conditions, environmental features, and urban infrastructure.

Understanding the factors that contribute to crime rates is essential for effective public policy and resource allocation. With the rise of data-driven approaches in public policy, there is an increasing interest in using data science and machine learning techniques to analyze crime patterns and develop predictive models.

In this project, we focus on crime in the Tucson area and use publicly available datasets from the City of Tucson Data Hub to investigate potential contributing factors to crime rates. More specifically, we aim to explore correlations between crime (m) the following features: (a) Neighborhood income, (b) Proximity to streetlights. There are several reasons for this focus. First, Tucson is famous for having low streetlight density, due to astronomy and light pollution concerns. We are interested in understanding how this affects crime rates and wether this causes higher crime rates in areas with fewer streetlights. Second, Tucson has many neighborhoods with varying income levels, which can influence crime rates. Figure 1 shows the top and bottom 20 neighborhoods by arrest rate, and Figure 2 shows the top and bottom 20 neighborhoods by median household income. Interestingly, we observe a significant overlap between the neighborhoods with the highest arrest rates and the lowest median household incomes. This suggests that there may be a correlation between income levels and

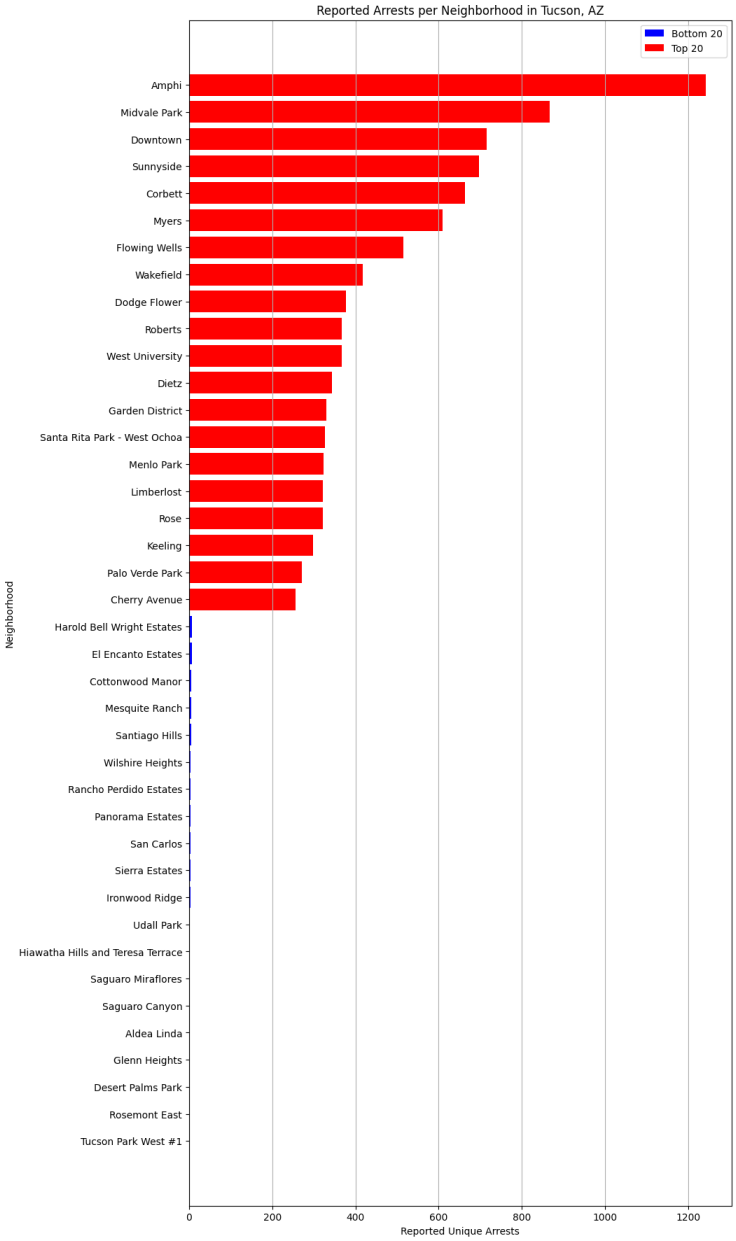

Figure 1: Top and bottom 20 neighborhoods by arrest rate

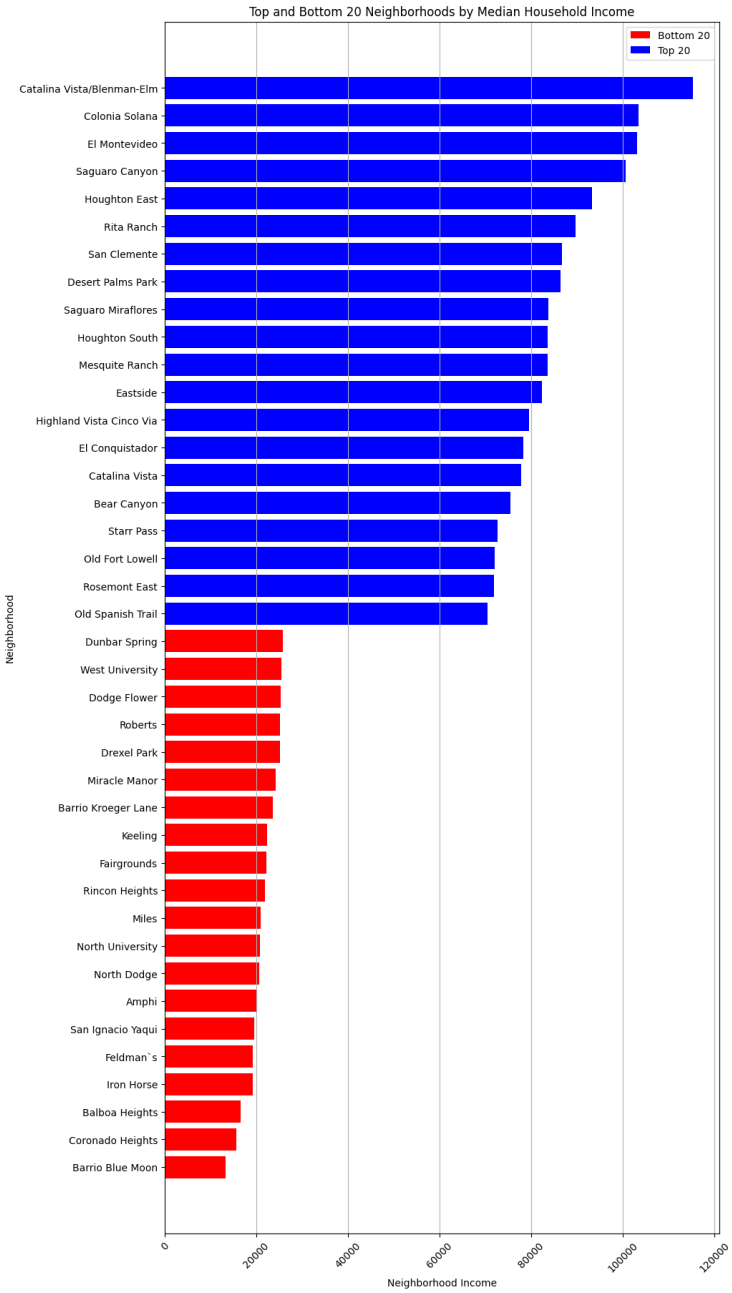

Figure 2: Top and bottom 20 neighborhoods by median household income

crime rates in Tucson. Additionally, given that streetlight is only relevant at night, we separate our analysis into daytime and nighttime crime rates. Surprisingly, our initial analysis shows that the areas with higher household income tend to have a higher density of streetlights, making it more interesting to investigate the relationship between streetlight density and crime rates.

We don't work with any human subjects in this project and do not make any ethical claims. However, given that we are building a predictive model for crime rates, the readers should be aware of the ethical implications of using such models in real-world applications. We acknowledge that predictive modeling can be a double-edged sword, as it can help improve public safety but also has the potential to reinforce existing biases and discriminatory practices.

2 Related Work

Previous work has explored how environmental and socioeconomic features affect crime patterns. Welsh and Farrington [2008] conducted a systematic review and found that improved street lighting was associated with a 20% average reduction in crime. Interestingly, this reduction was observed during both nighttime and daytime, when streetlights are not typically in use. This suggests that the presence of streetlights may have a broader impact on crime rates than just increasing visibility.

Recent experimental work by Chalfin et al. [2022] provides causal evidence supporting these observational findings. In a randomized trial in New York City public housing, the addition of streetlights led to a 36% reduction in outdoor nighttime index crimes.

This is more work in predictive policing focuses on using historical crime data to forecast future hotspots. Mohler et al. [2015] introduced a self-exciting point process model for predicting crime in space and time. Their method outperformed traditional analyst-led deployments in reducing crime, offering a compelling case for automated geospatial risk modeling.

Beyond infrastructure and historical data, neighborhood-level socioeconomic factors are widely known to correlate with crime. Redmond and Baveja [2002] introduced the *Communities and Crime* dataset, which links crime statistics to over 100 community-level features such as income, education, and racial composition. Many follow-up studies have found strong associations between indicators of disadvantage and crime rates.

Our work builds on these findings by modeling crime in Tucson using streetlight data, neighborhood income, and police-reported incidents. We contribute by linking infrastructure and socioeconomic indicators in a machine learning framework to better understand spatial variation in crime risk.

3 Methods

3.1 Data Collection and Preprocessing

We collected multiple datasets from the City of Tucson Data Hub, including:

- Police arrest records [City of Tucson, 2024]
- Neighborhood-level median household income [City of Tucson GIS IT, 2019]
- Streetlight locations [City of Tucson GIS IT, 2023]
- Parcel-level land information [City of Tucson GIS IT and Pima County, 2021]

We went through the following preprocessing steps:

- Coordinates were mapped to 1km^2 spatial grid cells.
- Arrest timestamps were parsed to extract hour-of-day, allowing for classification into `daytime` and `nighttime` incidents.
- Neighborhood income was merged into each grid cell based on polygon intersection.
- A `streetlights_per_sqkm` feature was computed by counting the number of streetlights within each grid.
- Crime type filtering was applied to retain only property-related incidents (theft, burglary, vandalism, etc.).

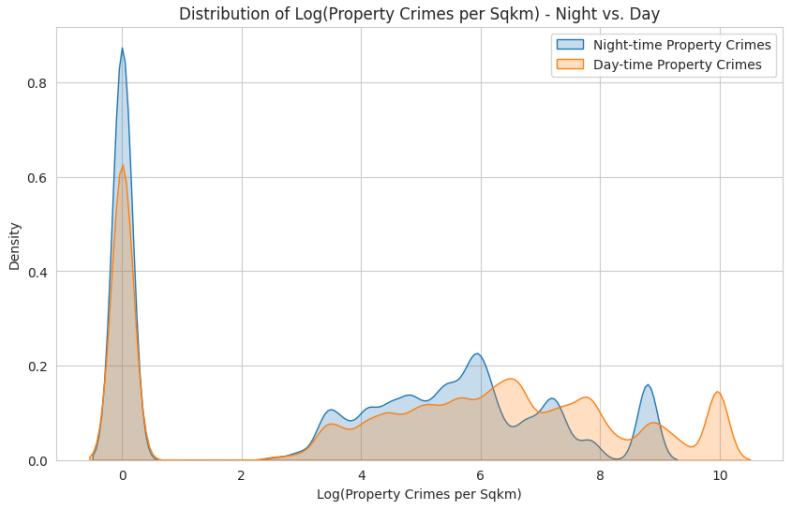

Figure 3: KDE distributions comparing nighttime and daytime property crimes

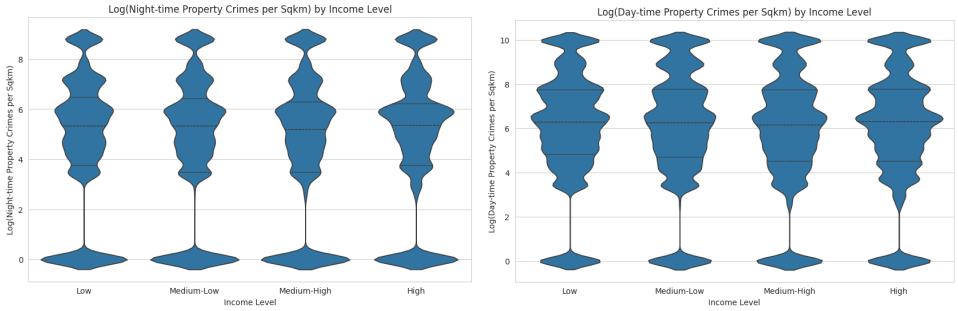

Figure 4: Violin plots by income group for night and day property crimes separately

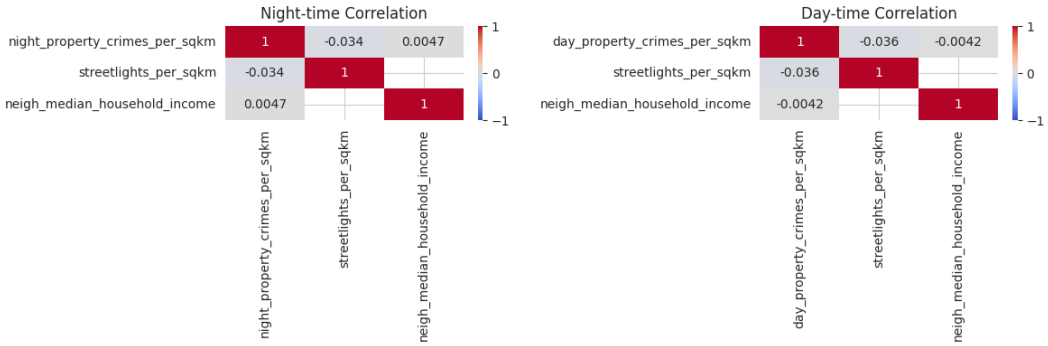

Figure 5: Heatmaps for nighttime and daytime correlations

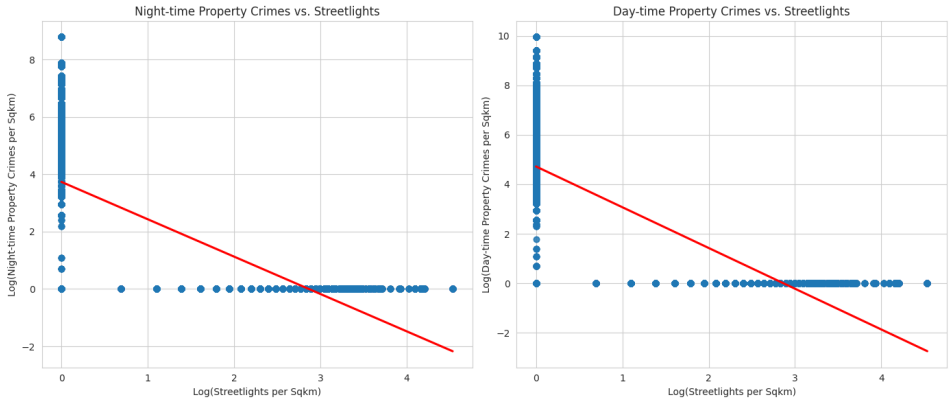

Figure 6: Regression plots: Streetlights vs. Property Crimes (Night vs Day)

3.2 Data Exploration

We first compared crime distributions between day and night, and across income brackets. KDE plots (Figure 3) show that nighttime crimes exhibit higher variance and heavier tails. Violin plots (Figure 4) visualize this stratified by income groups, revealing that lower-income neighborhoods experience higher and more variable nighttime property crime rates, while income has a weaker influence on daytime crime patterns.

Figure 5 shows correlation heatmaps for nighttime and daytime property crimes. We observe a negative correlation between streetlight density and property crime rates, both at night and during the day. Interestingly, the correlation between neighborhood income and property crime rates is positive during the night and negative during the day. This suggests that the relationship between income and crime rates is more complex than a simple linear correlation.

3.3 Modeling Framework

We define two modeling tasks: regression and classification. First, we build a linear regression model to to predict the continuous value of property crime rates (log-transformed) based on streetlight density, neighborhood income, and parcel density. This allows us to understand the relationship between these socioeconomic and infrastructural features and crime rates.

Second, we build a Random Forest classifier to predict whether a grid cell is `High Risk` or `Low Risk` for property crime. We define `High Risk` as a grid cell with a property crime rate above the median, and `Low Risk` otherwise. We use the same input features as in the regression model. A Random Forest classifier is chosen for its robustness to overfitting and ability to capture non-linear relationships.

Lastly, to isolate and have a better understanding of the influence of neighborhood income—which as we shall see is the most important feature in our analysis—we also build a linear (and logarithmic) regression model to predict property crime rates based solely on neighborhood income.

3.4 Model Evaluation

We evaluate the regression model using Mean Squared Error (MSE) and R^2 metrics. For the classification model, we use confusion matrices and ROC curves to assess performance. We also compute precision, recall, and F1-score to evaluate the model's ability to correctly classify high-risk areas. In order to run our expreiments, we split the data into training and testing sets, using 80% of the data for training and 20% for testing.

4 Results

The linear regression analysis with all three features (streetlight density, neighborhood income, and parcel density) achieves a root mean square error (RMSE) of 6.94 in the nighttime model and 7.49 in

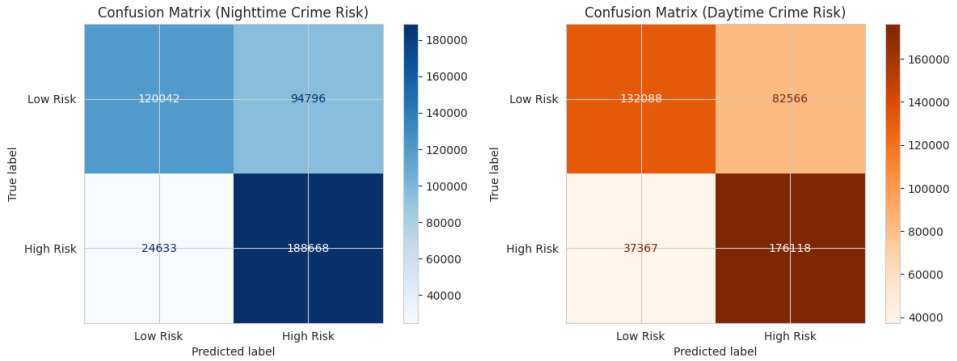

Figure 7: Confusion matrices for nighttime and daytime classification

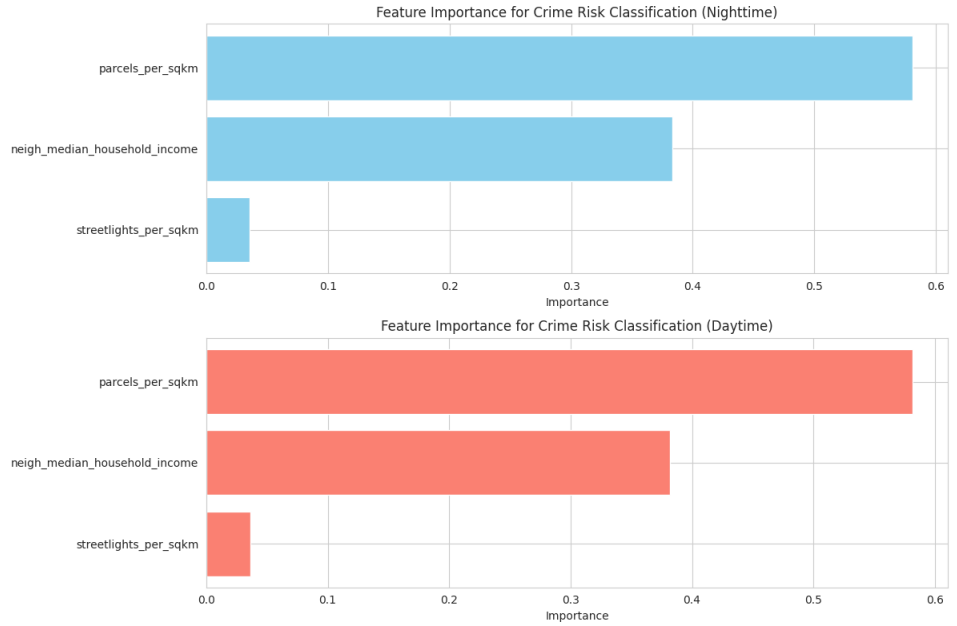

Figure 8: Feature importance for Crime Risk Classification

the daytime model. This indicates that the model is able to learn a better representation of the data at night than during the day. Given that we have three features, a visualization of the regression model is not straightforward.

In an ablation study, we investigated the influence of streetlight density in isolation. Figure 6 shows the relationship between streetlight density and property crime rates, both at night and during the day. This regression analysis shows that, perhaps surprisingly, there's no clear linear relationship between streetlight density and property crime rates during the day and night.

The results for classification, however, are more promising. Figure 7 shows the confusion matrices for nighttime and daytime classification tasks. We achieve a f-1 score of 0.72 in both nighttime and daytime classification tasks. The more interesting result is the feature importance analysis, as shown in Figure 8. We observe that the most important feature for both nighttime and daytime classification tasks is parcel density, followed by neighborhood income and streetlight density. Neighborhood income ihas a 38% importance in both nighttime and daytime classification tasks, while streetlight density has a 3.5% importance in daytime classification and 7.6% in nighttime classification. This suggests that streetlight density is not a strong predictor of crime rates as we initially expected. We also ran a more enhanced Random Forest classifier with better visualization and using more

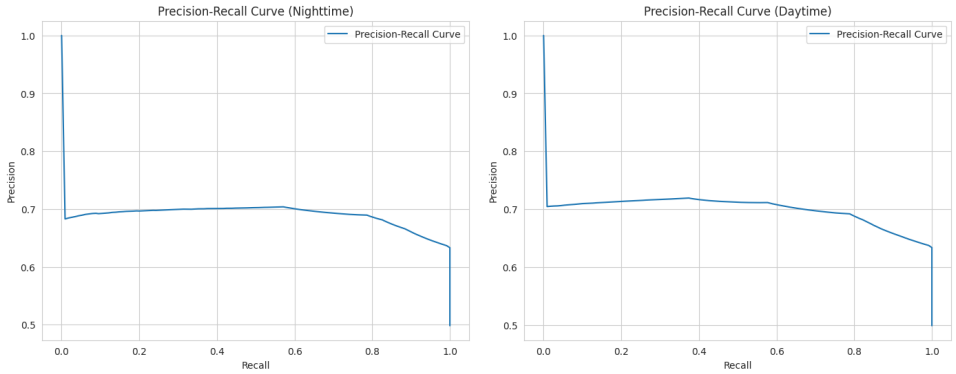

Figure 9: ROC curve for nighttime and daytime classification

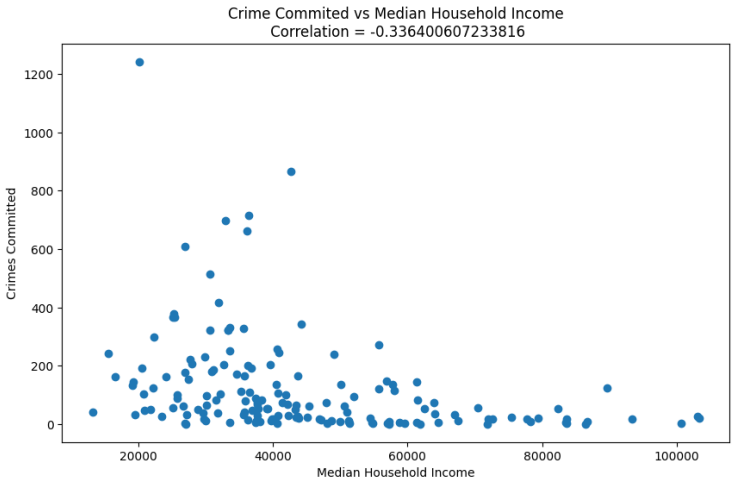

Figure 10: Crime commiteed vs median household income

acdvanced metrics. Figure 9 shows the ROC curves for nighttime and daytime classification tasks. R^2 scores are 0.25 for nighttime and 0.33 for daytime classification tasks.

Based on the presentend results, we realize that there is a significant correlation between neighborhood income and crime rates, as also shown in Figure 10. Therefore, we also ran a linear regression model to predict property crime rates based solely on neighborhood income. Figure 11 shows the linear regression model, which achieves an RMSE of 131.8 and an R^2 score of -0.03. We are able to observe a linear relationship between neighborhood income and property crime rates. We also ran a logarithmic regression model to predict property crime rates based solely on neighborhood income. Figure 12 shows the logarithmic regression model, which achieves an RMSE of 167.3 and an R^2 score of 0.11.

5 Conclusion

In this project, we examined how neighborhood income and streetlight density correlate with property crime rates in Tucson, Arizona. By integrating multiple public datasets and aggregating spatial information into 1km^2 grid cells, we were able to construct and evaluate predictive models using both regression and classification approaches.

Our analysis revealed that neighborhood income is consistently correlated with crime rates, with lower-income areas experiencing more property crime. While we hypothesized that streetlight density would also serve as a strong predictor—especially for nighttime crime—the data did not support a strong or consistent relationship. Regression analysis showed no clear linear trend between streetlights

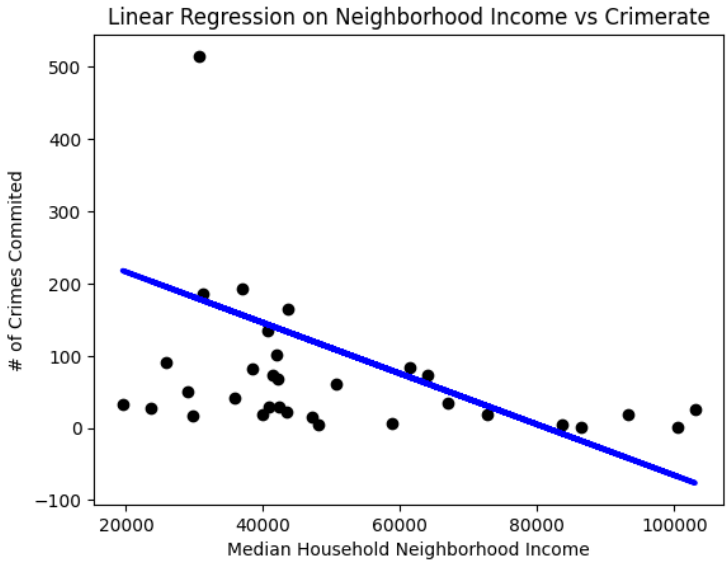

Figure 11: Linear regression model for neighborhood income and crime rate

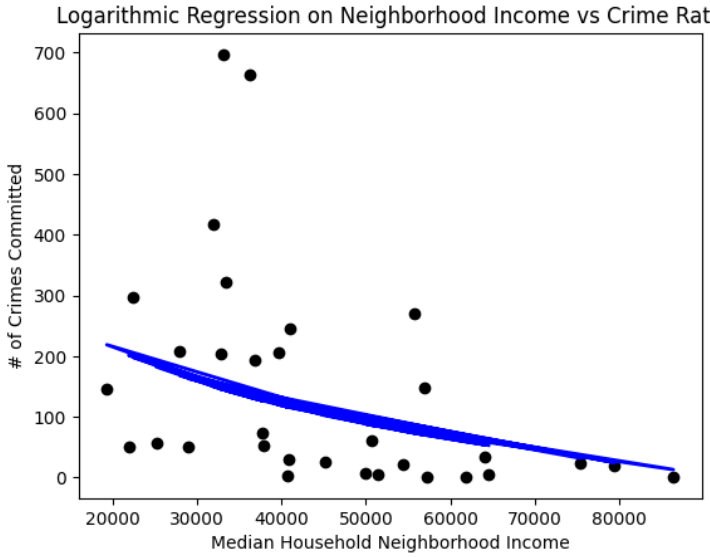

Figure 12: Logarithmic regression model for neighborhood income and crime rate

and crime, and feature importance analysis confirmed that streetlight density contributed relatively little to prediction performance compared to parcel and income density.

Classification models achieved reasonable F1-scores (0.72) and demonstrated that parcel density and neighborhood income were the most influential predictors of high-risk areas. Interestingly, the performance of models did not differ substantially between daytime and nighttime classifications, but nighttime crime rates showed slightly clearer patterns in both distribution and model fit.

Future work could explore richer temporal features, incorporate lighting outage data, and evaluate additional infrastructural indicators such as foot traffic or police presence. Moreover, incorporating community engagement and policy feedback loops can help ensure that predictive models support equitable and just outcomes.

References

Aaron Chalfin, Benjamin Hansen, Jason Lerner, and Lucie Parker. Reducing Crime Through Environmental Design: Evidence from a Randomized Experiment of Street Lighting in New York City. *Journal of Quantitative Criminology*, 38:127–157, 2022. doi: 10.1007/s10940-020-09490-6.

City of Tucson. Tucson Police Reported Crimes. `https://gisdata.tucsonaz.gov/datasets/tucson-police-reported-crimes/explore`, 2024. URL `https://gisdata.tucsonaz.gov/datasets/tucson-police-reported-crimes/explore`. Open dataset from the Tucson Police Department covering reported crimes from January 2017 to present. Updated monthly. Accessed May 2024.

City of Tucson GIS IT. Neighborhood Income. `https://gisdata.tucsonaz.gov/datasets/59f033d07eae41b0bdc21db87375d721_0`, 2019. URL `https://gisdata.tucsonaz.gov/datasets/59f033d07eae41b0bdc21db87375d721_0`. Neighborhood-level income data aggregated from block-level Esri 2019 demographic estimates (2010–2019). Accessed May 2025.

City of Tucson GIS IT. Streetlights - City of Tucson - Open Data. `https://gisdata.tucsonaz.gov/datasets/09ed59b6aae2483aa1bd32837d4aa7e5_19`, 2023. URL `https://gisdata.tucsonaz.gov/datasets/09ed59b6aae2483aa1bd32837d4aa7e5_19`. Geospatial dataset showing the location of streetlights maintained by the City of Tucson Transportation Department. Updated nightly. Accessed May 2025.

City of Tucson GIS IT and Pima County. Parcels - Regional. `https://gisdata.tucsonaz.gov/datasets/f87f048f6e9f4655bbc93efef4d65d05_12`, 2021. URL `https://gisdata.tucsonaz.gov/datasets/f87f048f6e9f4655bbc93efef4d65d05_12`. Geospatial dataset of tax parcels, common areas, and private roadway parcels in Pima County. Deprecated but used for detailed location and valuation features. Accessed May 2025.

George O. Mohler, Martin B. Short, Stephen Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi, and P. Jeffrey Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015. doi: 10.1080/01621459.2015.1077710.

Michael A. Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141 (3):660–678, 2002.

Brandon C. Welsh and David P. Farrington. Effects of improved street lighting on crime: a systematic review. *Campbell Systematic Reviews*, 4(1):1–51, 2008. doi: 10.4073/csr.2008.13.