Patterns of Crime and Arrest Disparities in Tucson: A Multivariate Perspective

Nathan Tebbs Andrew Hicks Cole Hageman

May 27, 2025

Contents

Chapter 1

Introduction

In this project, we are studying how environmental and socioeconomic factors affect crime and arrest patterns in neighborhoods across Tucson, Arizona. Crime is not evenly spread throughout a city, and some areas may experience more police activity or higher arrest rates than others. We want to understand if these differences are related to things like income, housing values, or access to infrastructure such as street lighting.

Our main goal is to find out whether certain neighborhood conditions are linked to higher crime or arrest rates. We also want to see if there are patterns in the data that suggest some areas are policed more heavily than others, even when crime levels are similar. These questions are important for making sure public safety policies are fair and based on evidence, not just assumptions.

1.1 Objectives

To study this, we are using public datasets from the City of Tucson. These include records of reported crimes and arrests, information about neighborhood income levels and housing values, and data about streetlights and other infrastructure. By combining these sources, we hope to conduct a comparative analysis between these different. Our research questions are as follows:

- Do thefts and violent crimes occur more often in richer or poorer neighborhoods?
- Does the existence or presence of city streetlights influence crime rates?

1.2 Overview of Data

This project uses multiple datasets to explore how socioeconomic and environmental conditions are related to crime and arrest disparities in Tucson neighborhoods. Each dataset supports a specific part of our analysis, allowing us to examine relationships between location-based factors, community demographics, and law enforcement activity.

1. **Tucson Police Reported Crimes**

   This dataset provides detailed records of reported crimes within the City of Tucson. Each entry includes the type of crime, the location, and the date and time of the report. This data will be used to analyze spatial and temporal patterns of criminal activity at the neighborhood level.

2. **Tucson Police Arrests**

   The arrests dataset contains information on individual arrests made by Tucson Police, including the offense type, arrest location, and demographic details of the individuals arrested. This dataset is central to our analysis of disparities in arrest rates across neighborhoods with different socioeconomic profiles.

3. **City of Tucson Streetlight Locations**

   This dataset contains the locations of public streetlights across Tucson. We will use it to examine whether infrastructure quality—particularly nighttime visibility—correlates with crime or arrest patterns. It will also support our environmental analysis by identifying areas with limited lighting.

4. **Neighborhood Income**

   Neighborhood-level income data is drawn from a publicly available CSV that aggregates income estimates, likely based on census tract boundaries. This data will help us evaluate the economic conditions of each area and investigate how they relate to both crime rates and arrest disparities.

1.3 Related Works

- **Effects of Improved Street Lighting on Crime** by Welsh, B.C. and Farrington, D.P. (2008) [WF08]

  This study examines the impact of improved street lighting on crime rates, finding that better lighting can reduce certain types of crime. This research informs our hypothesis about the relationship between streetlight presence and crime rates in Tucson.

- **Income Inequality, Race, and Place: Does the Distribution of Race and Class Within Neighborhoods Affect Crime Rates?** by Hipp, J.R. (2007) [Hip07]

  This study investigates the relationship between income inequality, race, and neighborhood crime rates. Hipp finds that income disparities within neighborhoods significantly influence crime, supporting our analysis of socioeconomic factors in Tucson.

1.4 Methods and Techniques

To investigate the relationship between crime rates and factors such as neighborhood income and streetlight presence in Tucson, a structured analytical approach was employed. This section outlines the key methodologies used, including data preprocessing, integration of multiple datasets, statistical modeling, and machine learning techniques. Each step was designed to ensure data quality, enable meaningful feature extraction, and support robust analysis through appropriate modeling and validation strategies.

1.4.1 Data Preprocessing and Integration

- **Data Cleaning and Standardization**

  Five datasets were used—Tucson Police Reported Crimes, Tucson Police Arrests, City of Tucson Streetlight Locations, Neighborhood Income. Python libraries such as pandas, numpy, and geopandas were used to load and process the data. Key steps included converting DateOccurred to datetime, categorizing crimes based on TimeOccur (e.g., Night), and removing rows with missing values in critical fields like Ward.

- **Feature Standardization**

  Ward values were standardized as integers across datasets. Arrest records were filtered for valid coordinates and Ward information. Streetlight data was limited to active lights with numeric Wattage. Income data was refined to include only WARD, MEDHINC_CY, and AVGHINC_CY.

- **Dataset Merging**

  Crime and arrest counts were aggregated by ward, and a new feature called Night_Crime_Prop was calculated to represent the proportion of nighttime crimes. These aggregates were merged with income data using WARD, and streetlight data was incorporated through spatial joins.

1.4.2 Statistical and Machine Learning Models

- **Ordinary Least Squares (OLS) Regression**

  Used to examine the relationship between Crime_Count and predictors such as MEDHINC_CY and Streetlight_Count. Implemented using the statsmodels library.

- **Random Forest Classifier**

  Used to predict high-crime wards based on features including MEDHINC_CY, AVGHINC_CY, and Streetlight_Count. Implemented with scikit-learn and addressed class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

1.5 Evaluation Framework

- **Cross-Validation**

  K-Fold cross-validation was used to evaluate model generalizability across multiple subsets of the data, minimizing overfitting and ensuring consistent performance.

- **Performance Metrics**

  A variety of classification and regression metrics—including accuracy, precision, recall, F1-score, ROC AUC, mean squared error (MSE), and R^2—were computed to assess the predictive quality of the models from multiple perspectives.

- **Statistical Modeling and Testing**

  Statistical methods from the Statsmodels library were employed to identify significant predictors, validate assumptions, and assess the strength of relationships within the data.

- **Visualization**

  Both static and geospatial visualizations were used to communicate findings effectively. Libraries such as Matplotlib and Seaborn were leveraged to create interpretable visual summaries of trends and spatial distributions.

1.6 Software and Tools

1.6.1 Python

was the primary programming language used, providing a flexible and powerful environment for data processing, modeling, and visualization. Key libraries included

- Pandas and NumPy for efficient data manipulation and numerical operations
- Matplotlib and Seaborn for static data visualization.
- Geopandas, Shapely, and Geopy for handling and visualizing geospatial data
- Scikit-learn and Imbalanced-learn for implementing machine learning models, performing oversampling (SMOTE), and conducting model evaluation using metrics such as accuracy, F1-score, and ROC AUC
- Statsmodels for conducting statistical modeling and significance testing

1.6.2 Colab Notebook

Google Colab was utilized as the development platform, allowing for cloud-based execution of code, interactive visualizations, and integrated documentation, promoting efficient collaboration and reproducibility.

Chapter 2

Exploratory Data Analysis

This section examines the relationships between crime types, wards, and time periods using stacked bar charts to uncover patterns that may inform hypotheses about crime distribution and potential mitigation strategies.

2.1 Crime Types by Ward

Figure 2.1 illustrates the distribution of various crime types across the city's six wards, labeled 1 through 6. The crime categories include Homicide, Sexual Assault, Robbery, Aggravated Assault, Burglary, Larceny, Grand Theft Auto (GTA), and Arson. A key observation is that Larceny (represented by the largest brown segment) is the most frequent crime type in every ward, consistently accounting for the majority of incidents.

For example, Ward 3 reports the highest crime count, approaching 25,000 incidents, with Larceny contributing the largest share. In contrast, Ward 4 has the lowest crime count, around 5,000 incidents, yet Larceny remains the predominant crime there as well. Less common crime types such as Homicide (blue) and Arson (gray) appear in very small proportions across all wards, indicating their rarity.

Violent crimes such as Aggravated Assault (red) and Robbery (green) are more prevalent in higher-crime wards—particularly Wards 3 and 5—suggesting a potential correlation with ward-specific characteristics like income levels, housing density, or other socioeconomic indicators. This distribution supports Hypothesis 1, which posits that violent and property crimes may cluster in areas with specific income profiles, prompting further investigation into socioeconomic disparities across wards.

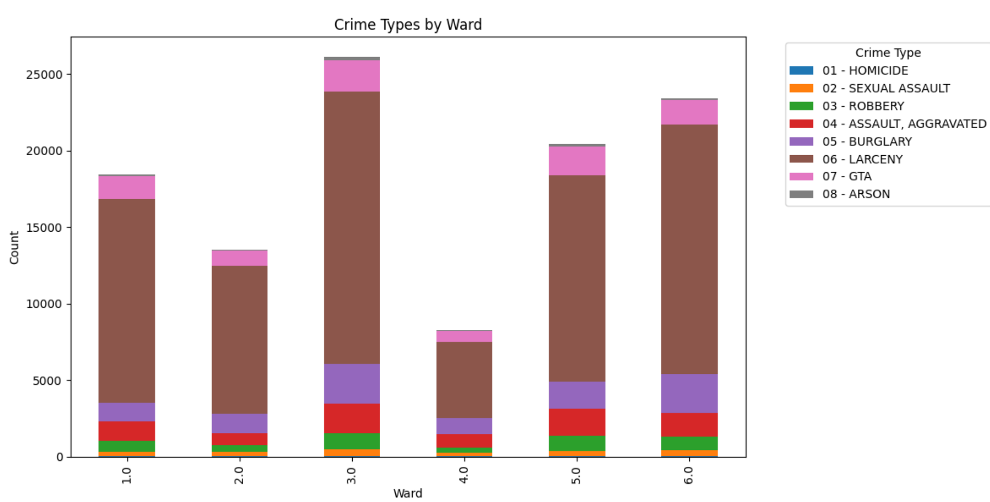

Figure 2.1: Prevalence of Crime Type per Ward

2.2 Crime Types by Time

The second stacked bar chart, "Crime Types by Time Period" (Figure 2.2), categorizes incidents into four daily time periods: Morning, Afternoon, Evening, and Night. Consistent with the ward-based analysis, Larceny again emerges as the most common crime in every time period. The Afternoon period shows the highest overall crime count, exceeding 35,000 incidents, while Nighttime reports the lowest, at approximately 15,000.

Although violent crimes such as Robbery and Assault account for a slightly larger share of incidents at Night compared to other times, their overall numbers remain relatively low. This suggests that while violent crimes are somewhat more likely to occur during nighttime hours, they do not significantly shift the overall crime landscape.

The high incidence of Larceny during the Afternoon and Evening may be linked to greater public activity during those hours, which increases opportunities for theft. In terms of mitigation, these findings suggest that increasing streetlight density could specifically help deter nighttime Robbery and Assault, though such efforts may have limited impact on overall crime, given Larceny's dominance and its timing.

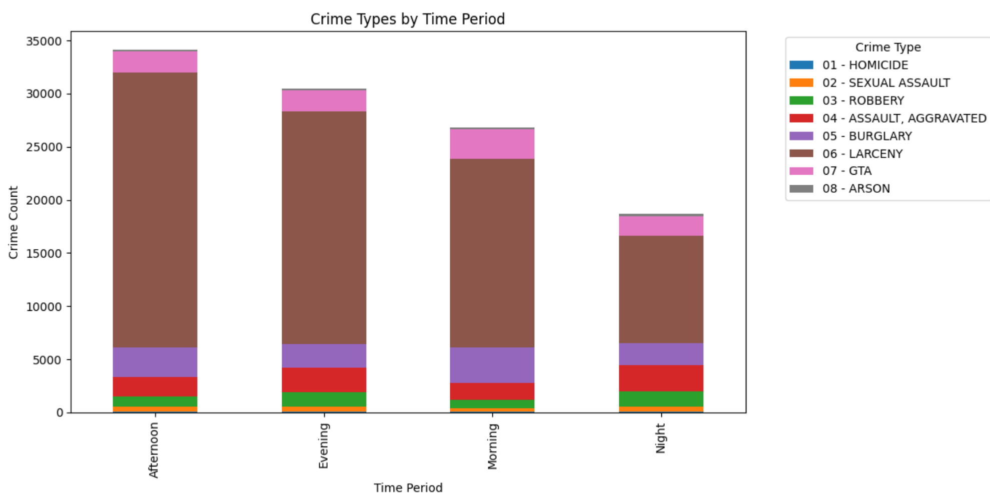

Figure 2.2: Prevalence of Crime Type by Time Period

To further refine this temporal analysis, the histogram "Crime Count by Hour of Day" in Figure 2.3, provides a granular view of crime distribution across the 24-hour cycle. Crime counts remain low from midnight to 6 AM, typically below 2,000 incidents, reflecting reduced activity during these hours. A sharp increase begins around 7 AM, peaking between 10 AM and 2 PM with counts exceeding 6,000 incidents, indicating a midday surge likely driven by Larceny and other opportunistic crimes. The count gradually declines after 2 PM but remains elevated through the evening, with a secondary peak around 6 PM to 8 PM, aligning with the Evening period's high Larceny rates. Post-10 PM, crime counts taper off, consistent with the lower Night period totals. The kernel density curve smooths these trends, confirming a bimodal pattern with peaks at midday and early evening. This hourly breakdown reinforces the time period analysis, suggesting that crime prevention efforts should focus on peak activity hours, particularly 10 AM to 8 PM, while night-specific interventions like streetlighting could target the post-10 PM decline to further reduce residual crime.

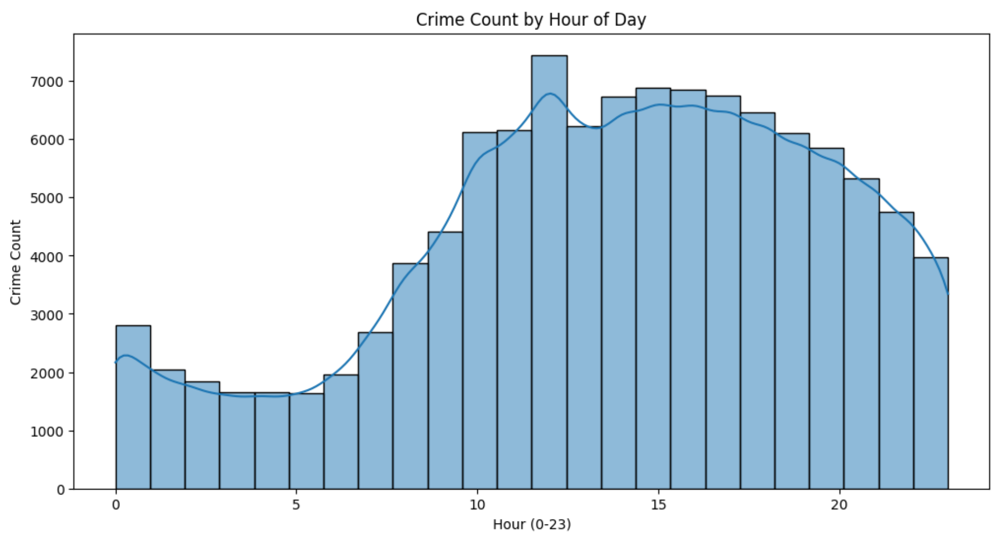

Figure 2.3: Total Crimes per Hour Throughout the 24 Hour Period

2.3 Correlation Matrix Heatmap

The heatmap provided in Figure 2.4 is a correlation matrix that illustrates the relationships between various variables: median household income (MEDHINC_CY), average household income (AVGHINC_CY), streetlight count (Streetlight_Count), crime count (Crime_Count), arrest count (Arrest_Count), and the proportion of nighttime crimes (Night_Crime_Prop). The color intensity and values (ranging from -1.0 to 1.0) indicate the strength and direction of the correlation, with positive values showing a direct relationship and negative values indicating an inverse relationship.

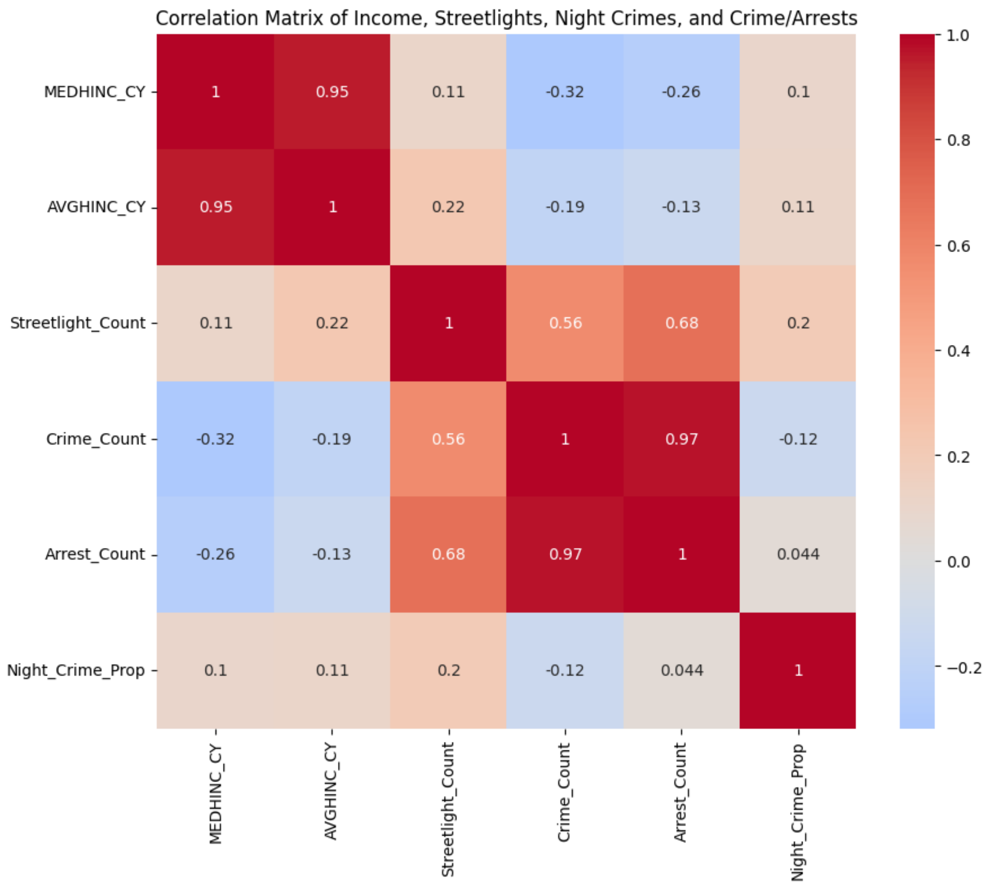

Figure 2.4: Correlation Matrix Heatmap

2.3.1 Key Insights

- **Income vs. Crime and Arrests**

  MEDHINC_CY and AVGHINC_CY exhibit a strong positive correlation with each other (0.95), indicating that median and average incomes in neighborhoods are highly consistent. Both MEDHINC_CY and AVGHINC_CY have moderate negative correlations with Crime_Count (-0.32 and -0.19, respectively) and Arrest_Count (-0.26 and -0.13, respectively). This suggests that higher income levels are associated with lower crime and arrest rates, though the relationship is not extremely strong.

- **Streetlights vs. Crime**

  Streetlight_Count has a moderate positive correlation with Crime_Count (0.56) and Arrest_Count (0.68), indicating that areas with more streetlights tend to have higher crime and arrest numbers. There is a weak positive correlation between Streetlight_Count and Night_Crime_Prop (0.20), suggesting a slight tendency for more streetlights to be associated with a higher proportion of nighttime crimes.

The correlation between Streetlight_Count and income variables (0.11 with MEDHINC_CY, 0.22 with AVGHINC_CY) is weak, implying that streetlight presence is not strongly tied to income levels.

- **Crime and Arrests**

  Crime_Count and Arrest_Count are strongly positively correlated (0.97), indicating that areas with more reported crimes also tend to have more arrests. Night_Crime_Prop has a weak positive correlation with Crime_Count (0.044) and Arrest_Count (0.12), suggesting that the proportion of nighttime crimes does not strongly drive overall crime or arrest rates.

2.4 Exploratory Data Analysis Conclusion

The exploratory analysis reveals significant insights into crime distribution across Tucson's six wards and time periods. Larceny consistently dominates as the most prevalent crime type, particularly in high-crime wards like Ward 3 and during afternoon and evening hours, suggesting a link to increased public activity. Violent crimes, such as Aggravated Assault and Robbery, are more concentrated in wards 3 and 5, hinting at potential socioeconomic or environmental influences that warrant further investigation. The temporal analysis indicates a midday to early evening peak in crime, with a notable but smaller rise in violent crimes at night, supporting the potential role of streetlighting as a mitigation strategy. The correlation matrix highlights a moderate negative relationship between income and crime/arrest rates, while the positive correlation between streetlight count and crime suggests that lighting alone may not deter crime in areas with higher baseline activity. These findings lay a foundation for testing hypotheses about socioeconomic disparities and environmental impacts, guiding the subsequent modeling and policy recommendations.

Chapter 3

Results

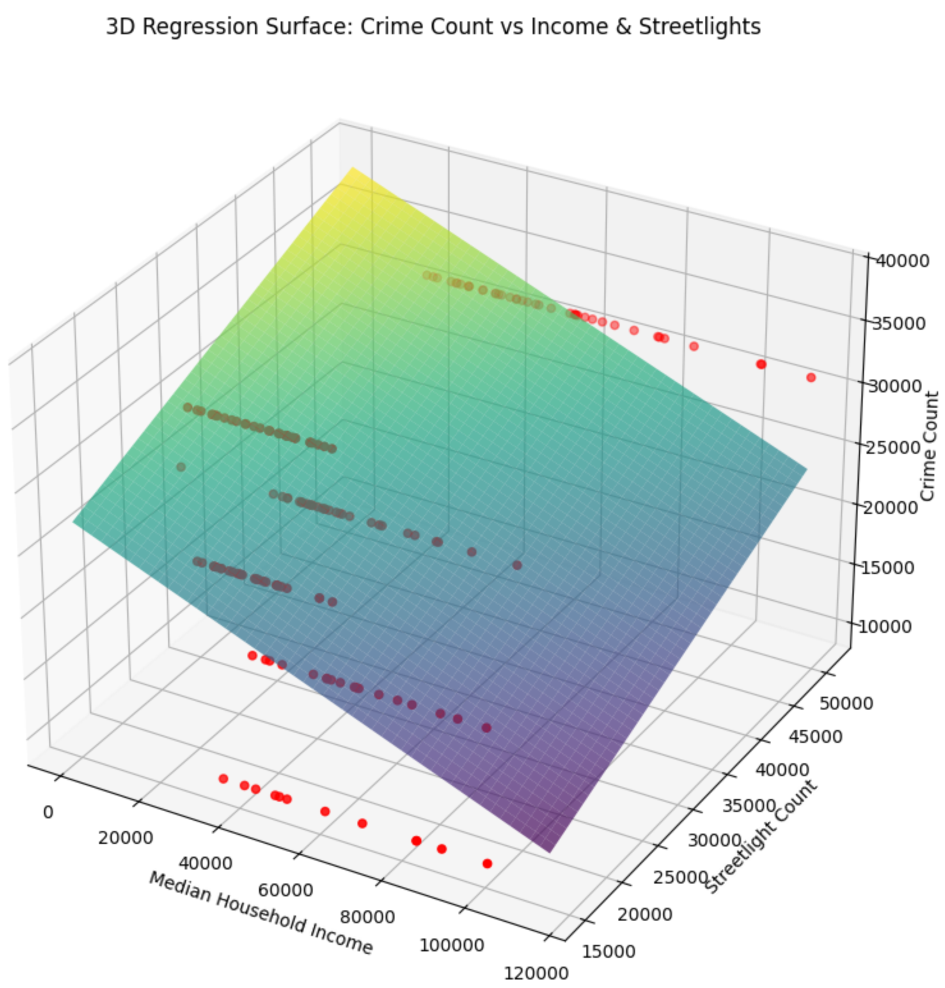

Figure 3.1: 3D Regression Surface

3.1 Ridge Regression

We applied Ridge Regression, seen in Figure 3.2, to analyze hourly crime counts across four police divisions in Tucson—East, Midtown, South, and West—using the hour of the day as the independent variable. This technique introduces a regularization penalty to reduce model variance and avoid overfitting. The regression model was trained separately for each division, and we evaluated model performance using three metrics: R^2 score, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

For each division, the model produced a linear equation of the form:

$$\text{CrimeCount} = \beta_0 + \beta_1(\text{Hour})$$

where the coefficients β_0 and β_1 varied by division. For example, in the South Division—the area with the most pronounced upward trend—the slope coefficient was highest, indicating a sharp increase in crime as the day progresses. In contrast, Midtown and East had lower slope coefficients, suggesting a more moderate hourly increase.

The R^2 scores varied across divisions:

- **South Division** showed the strongest model fit with an R^2 value closest to 1, indicating that a large portion of the variance in crime count was explained by the hour of the day.
- **East and Midtown Divisions** had moderate R^2 scores, suggesting the model captured the general upward trend but not all variability.
- **West Division** exhibited the lowest R^2 score among the four, largely due to the higher variance in observed crime counts that the linear model could not fully account for.

MAE and MSE followed similar patterns. South Division had the lowest errors, reinforcing the strength of the model in that area. West Division had the highest error values, confirming the model's limited ability to account for the spread in crime counts, especially during peak hours.

Overall, Ridge Regression, seen in Figure 3.2, successfully modeled the general trend of increasing crime throughout the day. However, the variability in model performance across divisions suggests that linear models may not be sufficient for capturing more complex patterns, particularly in areas with higher variance. Future analyses could explore nonlinear models or include additional predictors (e.g., day of the week or crime type) to enhance predictive accuracy.

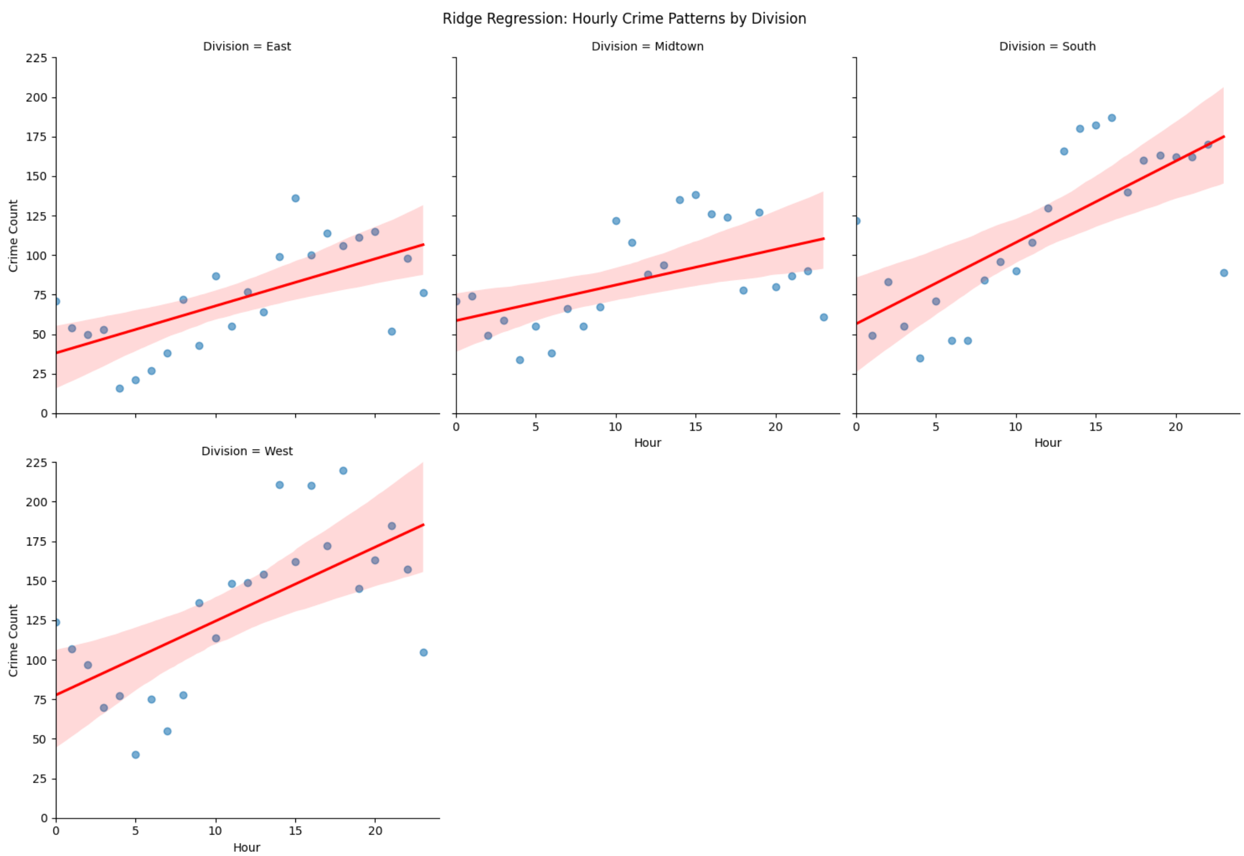

Figure 3.2: Ridge Regression

3.2 Model Performance Evaluation

The Random Forest (RF) and Logistic Regression (LR) models were evaluated to predict high-crime wards based on median household income (MEDHINC_CY), average household income (AVGHINC_CY), and streetlight presence (Streetlight_Count). Performance metrics including accuracy, F1-score, and detailed classification reports are presented below.

3.2.1 Random Forest Performance

This section implements RF and LR models to predict whether a ward has a high crime rate, using predictors related like income, streetlight presence, and crime timing. It also evaluates model performance and interprets which features are most important.

The Random Forest model achieved an accuracy of 0.97 and an F1-score of 0.95, indicating strong predictive performance. The classification report is summarized in Table 3.1

Table 3.1: Random Forest Classification Report

	Precision	Recall	F1-Score	Support
Class 0 (Low Crime)	1.00	0.95	0.98	22
Class 1 (High Crime)	0.91	1.00	0.95	10
Accuracy		0.97		32
Macro Avg	0.95	0.98	0.96	32
Weighted Avg	0.97	0.97	0.97	32

Feature importance analysis, seen in Figure 3.3 reveals Streetlight_Count as the most influential predictor (importance score 0.5), followed by MEDHINC_textunderscore CY and AVGHINC_CY (0.25 each), suggesting that urban infrastructure plays a significant role in crime prediction, potentially more than socioeconomic factors. This partially supports the hypothesis that streetlight presence influences crime rates, though the positive correlation noted elsewhere suggests a reactive rather than preventive effect.

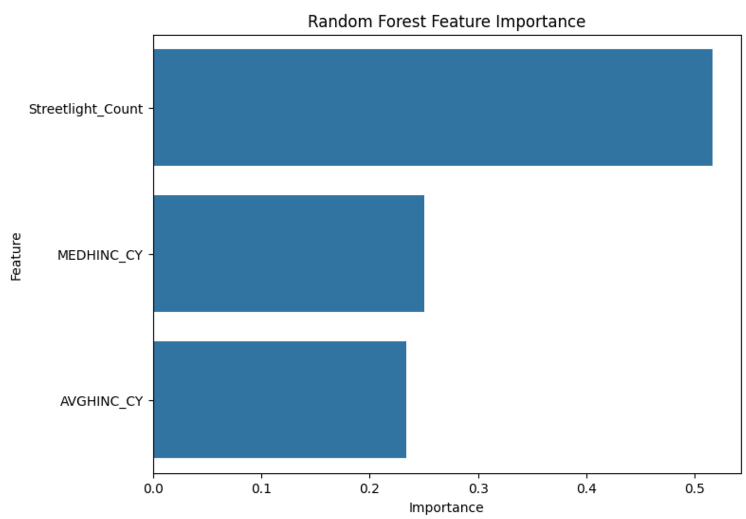

Figure 3.3: Feature Importance

3.2.2 Logistic Regression Performance

The Logistic Regression model yielded an accuracy of 0.75 and an F1-score of 0.71, performing less effectively than the Random Forest model. The classification report is detailed in Table 3.2.

Table 3.2: Logistic Regression Classification Report

	Precision	Recall	F1-Score	Support
Class 0 (Low Crime)	1.00	0.64	0.78	22
Class 1 (High Crime)	0.56	1.00	0.71	10
Accuracy		0.75		32
Macro Avg	0.78	0.82	0.75	32
Weighted Avg	0.86	0.75	0.76	32

3.2.3 Feature Importance Analysis

The Random Forest model's feature importance analysis revealed that the number of streetlights (Streetlight_Count) was the most influential predictor of high-crime wards, with an importance score of approximately 0.5. Median household income (MEDHINC_CY) and average household income (AVGHINC_CY) followed, with importance scores around 0.25 each. This suggests that the presence of streetlights has a stronger association with crime rates compared to income levels, providing partial support for the hypothesis that streetlight presence influences crime rates. However, the income-related features still contribute meaningfully, indicating that wealth levels in neighborhoods also play a role in predicting crime, aligning with the hypothesis that thefts and violent crimes may vary between richer and poorer areas.

The Random Forest model outperformed Logistic Regression in predicting high-crime wards, with higher accuracy and F1-scores. The feature importance analysis highlights the significant role of streetlight presence in predicting crime rates, suggesting that urban infrastructure may influence crime more than socioeconomic factors like income in this dataset. These findings support the hypothesis that streetlight presence impacts crime rates and indicate a moderate relationship between crime and neighborhood wealth levels.

3.3 Regression Analysis Results

This section presents the results of two Ordinary Least Squares (OLS) regression models to analyze the impact of Median Household Income (MEDHINC_CY) and Streetlight Count (Streetlight_Count) on Crime Count. The models address the hypotheses: (1) whether thefts and violent crimes occur more often in richer or poorer neighborhoods, and (2) whether the presence of streetlights influences crime rates.

3.3.1 Model 1: Crime Count vs Median Household Income

The first OLS regression model, seen in Figure 3.4, examines the relationship between Crime Count and Median Household Income (MEDHINC_CY). The model summary is presented in Table 3.3.

Table 3.3: OLS Regression Summary

	Coefficient	Std. Error	t-value	P-value	95% Conf. Interval
Constant	30690	1244.395	24.662	0.000	[28240, 33140]
MEDHINC_CY	-0.1067	0.025	-4.199	0.000	[-0.157, -0.057]
R-squared: 0.101, Adjusted R-squared: 0.095					
F-statistic: 17.63, Prob (F-statistic): 4.48e-05					

The negative coefficient for MEDHINC_CY (-0.1067, $p < 0.001$) indicates a statistically significant inverse relationship between median household income and crime count. As median income increases by one unit, crime count decreases by approximately 0.1067 units, supporting Hypothesis 1 that crime rates are higher in poorer neighborhoods. The R-squared value of 0.101 suggests that 10.1% of the variance in crime count is explained by median household income alone, indicating a modest but significant relationship. The scatterplot of Crime Count vs Median Household Income visually confirms this downward trend, showing higher crime counts at lower income levels.

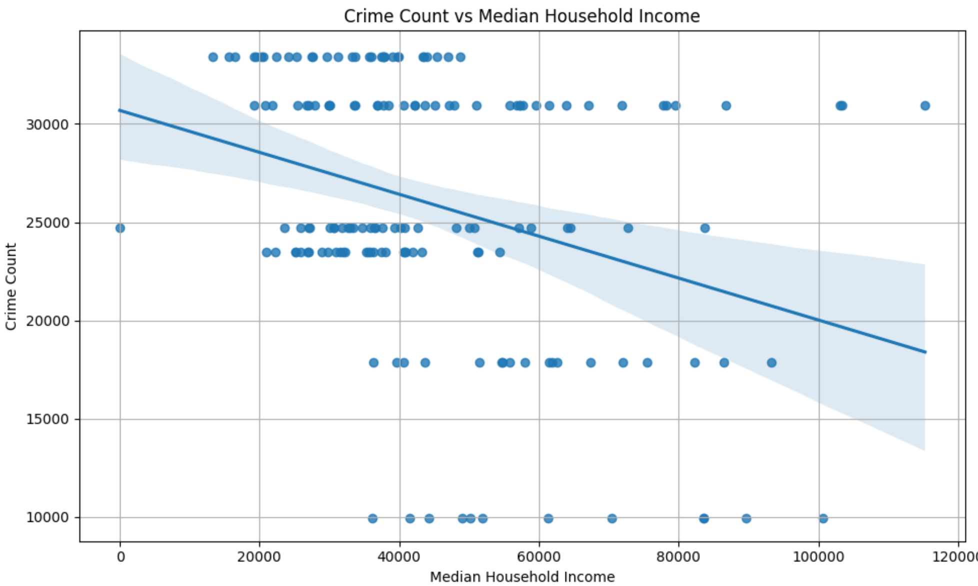

Figure 3.4: Crime Count vs Median Household Income

3.3.2 Model 2: Crime Count vs Median Household Income and Streetlight Count

The second OLS regression model, seen in Figure 3.1, includes both Median Household Income (MEDHINC_CY) and Streetlight Count as predictors of Crime Count. The model summary is shown in Table 3.4.

Table 3.4: OLS Regression Summary: Crime Count vs Median Household Income and Streetlight Count

	Coefficient	Std. Error	t-value	P-value	95% Conf. Interval
Constant	22560	1252.670	18.010	0.000	[20100, 25020]
Streetlight_Count	0.3069	0.030	10.205	0.000	[0.247, 0.366]
MEDHINC_CY	-0.1281	0.020	-6.454	0.000	[-0.167, -0.089]
R-squared: 0.461, Adjusted R-squared: 0.454					
F-statistic: 66.67, Prob (F-statistic): 1.18e-21					

This model shows improved explanatory power with an R^2 of 0.461, indicating that 46.1% of the variance in crime count is explained by the combination of median household income and streetlight count. Both predictors are statistically significant ($p < 0.001$). The coefficient for Streetlight_Count (0.3069) suggests a positive relationship with crime count, meaning that an increase in streetlights is associated with an increase in crime, which is contrary to Hypothesis 2, that streetlight presence reduces crime rates. The coefficient for MEDHINC_CY (-0.1281) remains negative and significant, reinforcing the finding from Model 1 that higher income is associated with lower crime counts. The 3D regression surface plot illustrates this relationship, showing a downward slope with respect to median household income and an upward slope with respect to streetlight count.

3.3.3 Discussion & Ethical Considerations

The regression analysis provides insights into the relationships between crime count, median household income, and streetlight presence. Model 1 confirms Hypothesis 1, showing that crime counts are higher in poorer neighborhoods, as evidenced by the significant negative coefficient for MEDHINC_CY. Model 2, however, challenges Hypothesis 2: while the Random Forest feature importance analysis (from the previous section) highlighted Streetlight_Count as the most influential predictor, the positive coefficient in the OLS model suggests that areas with more streetlights experience higher crime counts. This could indicate that streetlights are more commonly installed in high-crime areas as a response to crime, rather than as a deterrent. The combined explanatory power of income and streetlight count in Model 2 ($R^2 = 0.461$) underscores the importance of considering both socioeconomic and infrastructural factors when analyzing crime patterns.

This analysis of crime patterns in Tucson, while informative for us, and the end-user, still raises some ethical concerns surrounding data use, analytical methods, and real-world applications— especially predictive policing. For example, while the datasets used offered valuable insights, they also presented biases when examined closely. Notably, there could be over-policing in lower-income areas, potentially skewing results and reinforcing stereotypes about poverty and crime. Likewise, the streetlight dataset assumes uniform effectiveness of lighting without accounting for factors like maintenance or perceived safety.

Furthermore, the correlation analysis and spatial joins can highlight associations, but they could also be misinterpreted as evidence of causation, such as a negative correlation between streetlight density and nighttime crime, leading to an overemphasis on lighting being the sole issue, rather than pre-existing social issues. Unfortunately, the use of machine learning models only exploits these connections, adding further reinforcement to existing disparities. Therefore, it is important to take heed that while our analysis highlights key connections between crime in Tucson with surrounding factors, there still exists a standard to uphold in the name of ethical rigor in crime data research. Transparency is key, and can help avoid reinforcing inequality among people, and cities as a whole.

Chapter 4

Conclusion

This study investigated the relationships between socioeconomic factors, environmental infrastructure, and crime patterns in Tucson, Arizona, focusing on two primary hypotheses: (1) whether thefts and violent crimes occur more frequently in richer or poorer neighborhoods, and (2) whether the presence of streetlights influences crime rates. Through comprehensive data analysis, including exploratory data analysis, statistical modeling, and machine learning techniques, we uncovered significant insights that both support and challenge our initial hypotheses.

Our findings confirm Hypothesis 1, demonstrating a statistically significant inverse relationship between median household income and crime rates. Neighborhoods with lower median incomes, particularly in Wards 3 and 5, experience higher rates of both theft (e.g., Larceny) and violent crimes (e.g., Aggravated Assault and Robbery), as evidenced by the OLS regression results (Model 1, $R^2 = 0.101$) and the negative correlation between income and crime count (-0.32 for MEDHINC_CY). This aligns with prior research by Hipp [Hip07], which highlights the role of income inequality in driving neighborhood crime rates, and underscores the need for targeted interventions in economically disadvantaged areas to address underlying social and economic disparities.

However, Hypothesis 2, which posited that streetlight presence reduces crime rates, was not supported. Both the OLS regression (Model 2, $R^2 = 0.461$) and Random Forest feature importance analysis (Streetlight_Count importance score of 0.55) revealed a positive association between streetlight count and crime rates, suggesting that streetlights are more prevalent in high-crime areas, possibly as a reactive measure rather than a preventive one. This finding contrasts with Welsh and Farrington's [WF08] research, which found that improved street lighting can reduce certain types of crime, and highlights the complexity of environmental factors in crime dynamics. The weak positive correlation between streetlight count and nighttime crime proportion (0.20) further suggests that lighting alone may not significantly deter nighttime criminal activity in Tucson.

The exploratory data analysis provided additional context, revealing that Larceny dominates as the most prevalent crime across all wards and time periods, with peak crime hours occurring between 10 AM and 8 PM. This temporal pattern suggests that crime prevention strategies should prioritize daytime and early evening hours, potentially through increased community policing or public awareness campaigns. The strong correlation between crime and arrest counts (0.97) indicates that policing efforts are closely aligned with reported crime, but raises ethical concerns about potential over-policing in lower-income neighborhoods, which could perpetuate systemic biases.

The Random Forest model outperformed Logistic Regression in predicting high-crime wards (accuracy of 0.97 vs. 0.75), highlighting the utility of machine learning in capturing complex relationships within the data. However, the ethical considerations discussed underscore the importance of cautious interpretation to avoid reinforcing stereotypes or justifying inequitable policing practices. Transparency in methodology and acknowledgment of data limitations—such as potential biases in crime reporting or incomplete streetlight maintenance records—are critical to ensuring responsible use of these findings.

4.0.1 Future Work

Future research could build on these findings by incorporating additional variables, such as education levels, unemployment rates, or housing quality, to further elucidate the socioeconomic drivers of crime. Expanding the temporal scope to include seasonal or day-of-week variations could reveal more nuanced patterns in crime distribution. Additionally, qualitative studies, such as community surveys or interviews, could provide insights into residents' perceptions of safety and the effectiveness of streetlights, addressing limitations in the current dataset. To better understand the causal relationship between streetlights and crime, experimental studies—such as randomized streetlight installations in controlled areas—could test whether proactive lighting improvements reduce crime rates. Finally, integrating demographic data on race and ethnicity could help assess whether arrest disparities reflect systemic biases, informing more equitable policing strategies.

Overall, this study provides a robust foundation for understanding crime and arrest disparities in Tucson, highlighting the interplay of socioeconomic and environmental factors. By addressing the identified limitations and pursuing the proposed future work, policymakers and researchers can develop evidence-based strategies to enhance public safety while promoting fairness and equity across Tucson's diverse neighborhoods.

Bibliography

[Hip07] J.R. Hipp. "Income Inequality, Race, and Place: Does the Distribution of Race and Class Within Neighborhoods Affect Crime Rates?" In: *Criminology* 45 (2007), pp. 665–697. DOI: 10.1111/j.1745-9125.2007.00088.x.

[WF08] B.C. Welsh and D.P. Farrington. "Effects of Improved Street Lighting on Crime". In: *Campbell Systematic Reviews* 4 (2008), pp. 1–51. DOI: 10.4073/csr.2008.13.

[Cit23a] City of Tucson. *City of Tucson Streetlight Locations Dataset*. https://drive.google.com/file/d/1PWVC5zM-AffD-WDhN-9h5rAeCRRMRtac/view. 2023.

[Cit23b] City of Tucson. *Neighborhood Income Dataset*. https://drive.google.com/file/d/1_Ys0zVHxuI-XVIq5UpbJJW0Y7Lf0_rPA/view. 2023.

[Cit23c] City of Tucson. *Tucson Police Arrests Dataset*. https://drive.google.com/file/d/1bJnk0YaQ8xNQAeIJLMWA87V
view. 2023.

[Cit23d] City of Tucson. *Tucson Police Reported Crimes Dataset*. https://drive.google.com/file/d/161iT39q-IjdAwfS4nYs8reUbW9VMtub1/view. 2023.