

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

Nathan Tebbs

Andrew Hicks

Cole Hageman

May 7, 2025

Contents

1	Introduction	2
1.1	Objectives	3
1.2	Overview of Data	3
1.3	Related Works	4
1.4	Methods and Techniques	5
1.4.1	Data Preprocessing and Integration	5
1.4.2	Statistical and Machine Learning Models	5
1.5	Evaluation Framework	6
1.6	Software and Tools	6
1.6.1	Python	6
1.6.2	Colab Notebook	7
2	Exploratory Data Analysis	8
2.1	Crime Types by Ward	8
2.2	Crime Types by Time	9
2.3	Correlation Matrix Heatmap	11
3	Results	13
3.1	Ridge Regression	14

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

- **The Effects of Neighborhood Characteristics on Police Officers’ Decisions to Initiate Encounters” by Robin S. Engel, Michael R. Smith, and Robert E. Worden (2007) [1]**

The foundation of our project is supported by past research that explores the complex relationship between neighborhood-level conditions and disparities in crime and arrest patterns. Two key studies inform our approach to understanding how environmental and socioeconomic factors shape community-level justice outcomes.

- **The Effects of Neighborhood Characteristics on Police Officers’ Decisions to Initiate Encounters” by Robin S. Engel, Michael R. Smith, and Robert E. Worden (2007) [2]**

This study investigates whether crime is randomly distributed across urban space or concentrated in specific locations. Weisburd et al. analyze 16 years of crime data from Seattle and find that a small number of street segments account for a large and stable share of total crime, suggesting that crime hot spots are persistent over time. This research is important to our project because it highlights the spatial concentration of crime and the value of analyzing crime at a micro-geographic level. It supports the idea that certain areas consistently experience higher levels of crime, which informs our efforts to identify and interpret crime patterns across Tucson neighborhoods.

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 nine wards, labeled 1 through 9. 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 9—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.

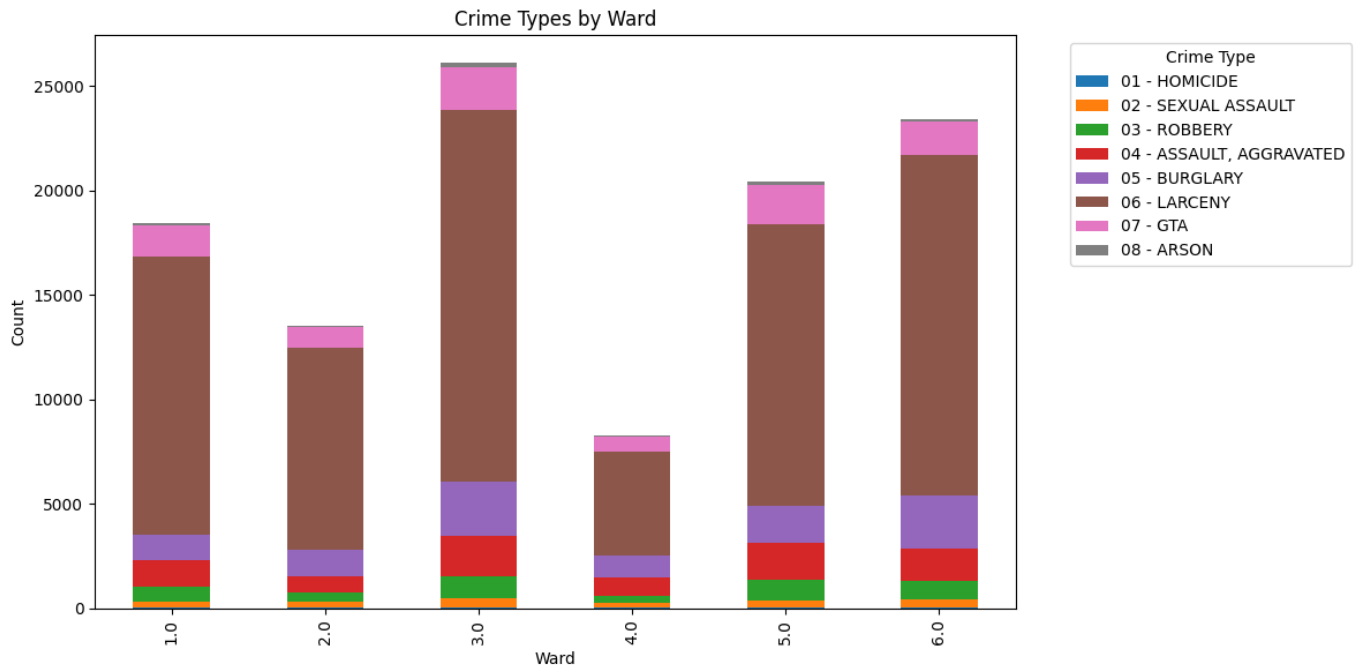


Figure 2.1: Prevalence of Crime Type per Ward

2.2 Crime Types by Time

The second stacked bar chart, “Crime Types by Time Period” (Figure 3.1), 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.

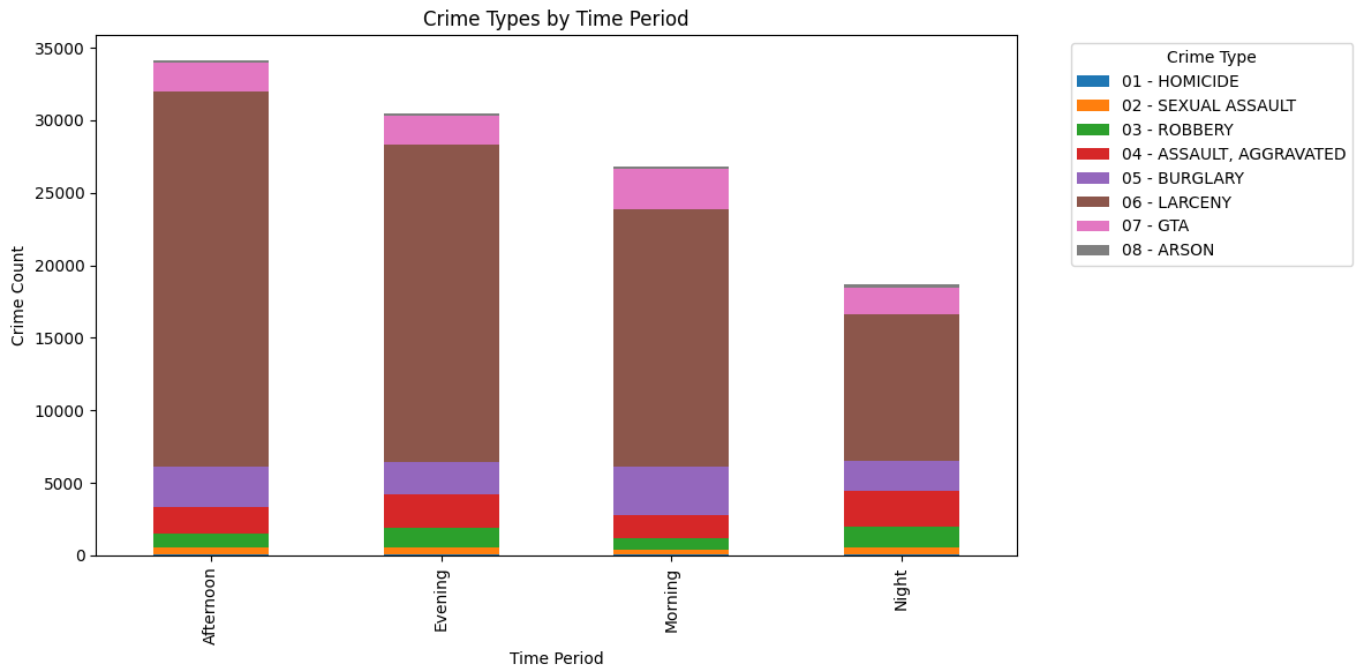


Figure 2.2: Prevalence of Crime Type per Ward

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.

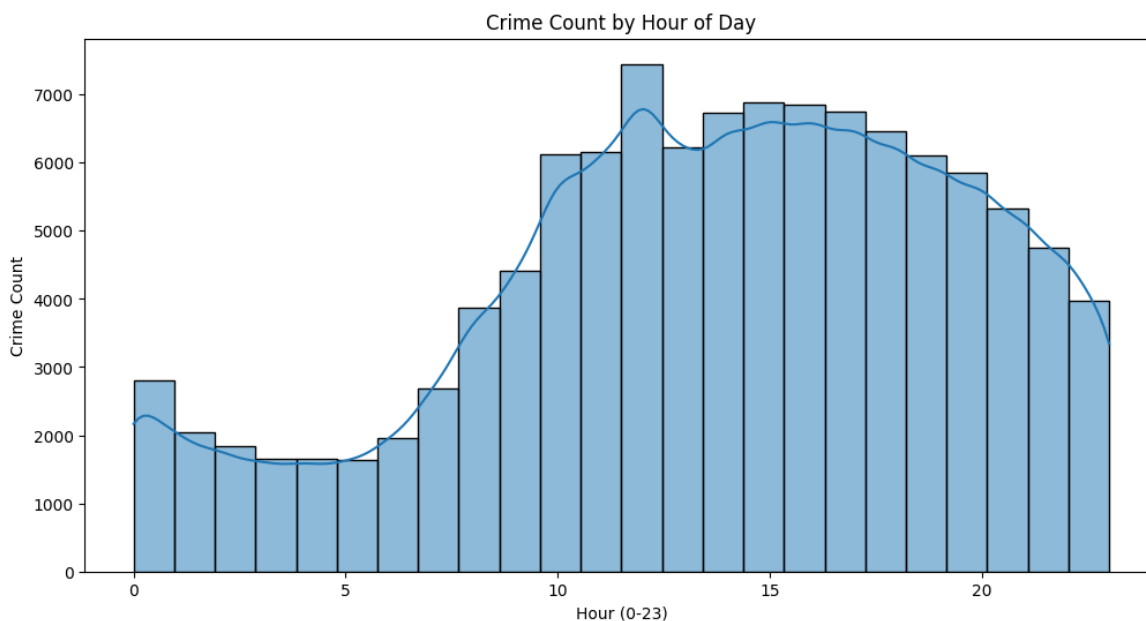


Figure 2.3: Correlation Matrix Heatmap

2.3 Correlation Matrix Heatmap

The heatmap in Figure 2.4 reveals relationships between median household income (MEDHINC_CY), average household income (AVGHINC_CY), streetlight count, crime count, arrest count, and the proportion of nighttime crimes (Night_Crime_Prop).

A strong positive correlation (0.95) exists between median and average income, indicating consistency in income measures across the dataset. Both income metrics show moderate negative correlations with crime count (-0.37 and -0.25, respectively), suggesting that higher income areas may experience lower crime rates, though the relationship is not overwhelming.

Streetlight count exhibits a weak negative correlation with crime count (-0.37), hinting at a potential deterrent effect, but this is tempered by its positive correlation with the proportion of nighttime crimes (0.034), possibly indicating that areas with more streetlights have a higher relative share of night crimes despite lower overall counts.

The proportion of nighttime crimes shows a weak positive correlation with income (0.14 and 0.16) and a moderate positive correlation with crime count (0.034), suggesting that night crimes may be slightly more prevalent in higher-crime, higher-income areas. These findings support further exploration into how streetlighting and income levels influence crime patterns, particularly at night.

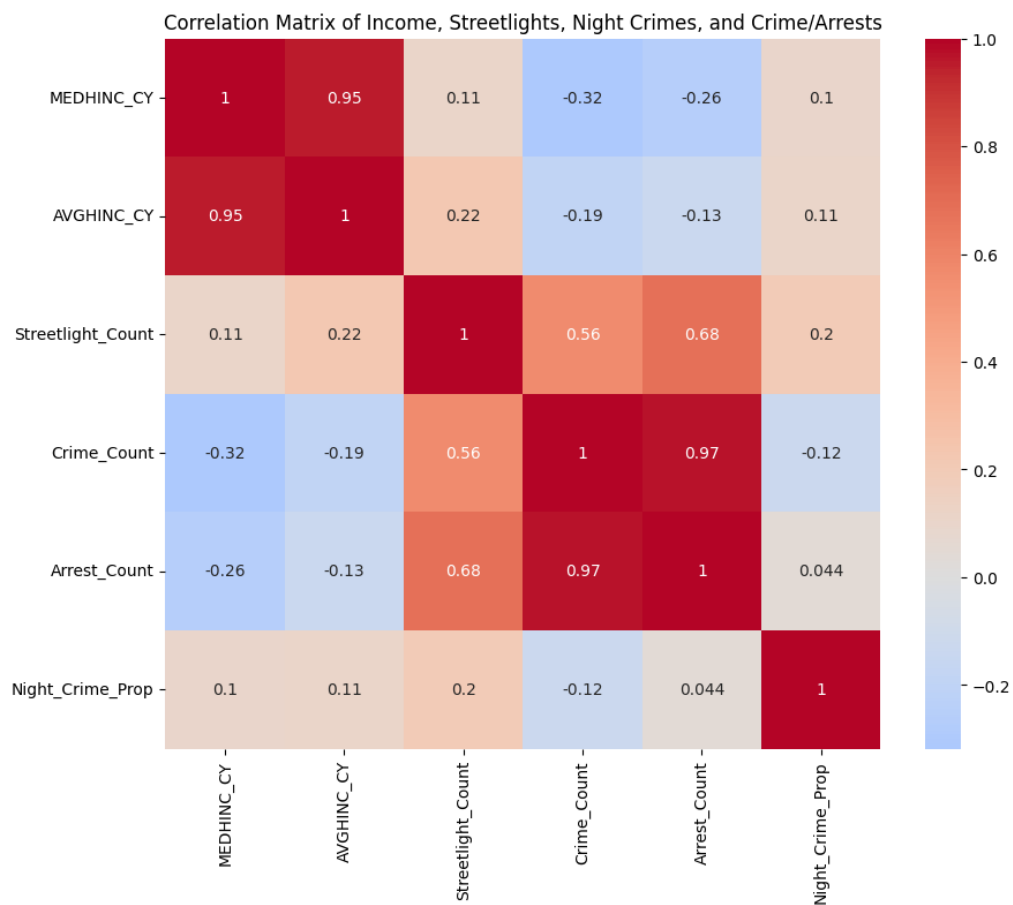


Figure 2.4: Correlation Matrix Heatmap

Chapter 3

Results

3D Regression Surface: Crime Count vs Income & Streetlights

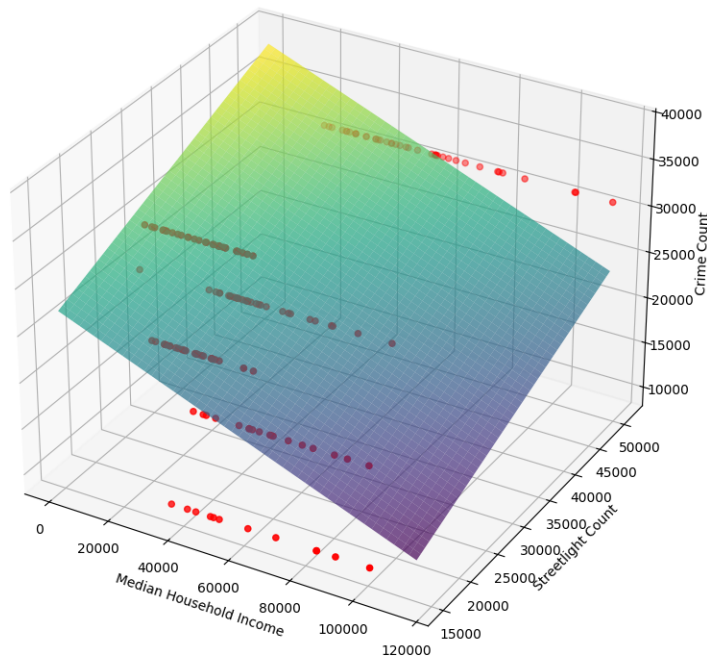


Figure 3.1: 3D Regression Surface

3.1 Ridge Regression

We applied Ridge Regression 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:

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

test test 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 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.

Bibliography

- [1] Welsh, B.C. and Farrington, D.P. (2008), Effects of Improved Street Lighting on Crime. Campbell Systematic Reviews, 4: 1-51.
<https://doi.org/10.4073/csr.2008.13>
- [2] HIPP, J.R. (2007), INCOME INEQUALITY, RACE, AND PLACE: DOES THE DISTRIBUTION OF RACE AND CLASS WITHIN NEIGHBORHOODS AFFECT CRIME RATES?*. Criminology, 45: 665-697.
<https://doi.org/10.1111/j.1745-9125.2007.00088.x>