Final Project - Apurbo Barua (netid: apurbobarua) & Muhammad Asifur Rahman (netid: asifrahman)

Course: CSc 380 - Principles of Data Science

Professor: Cesim Erten

Analysis Introduction: Tucson Crime Analysis

Primary Hypothesis: Low-income neighborhoods show higher arrest activity during late-night hours compared to high-income neighborhoods.

Elaborated Explanation: This hypothesis examines the correlation between neighborhood income levels and arrest activities during specific time periods in the city of Tucson, Arizona. It posits that late-night hours (e.g., 12 AM to 6 AM) might see a disproportionately high number of arrests in low-income neighborhoods compared to high-income neighborhoods. We think that Tucson's unique socio-economic landscape, coupled with the nature of its crime patterns, provides a compelling basis for exploring this relationship.

Rationale and Significance:

1. **Socioeconomic Disparities**: Low-income neighborhoods often face challenges like limited resources, higher unemployment rates, and increased stressors, which can contribute to higher crime rates. Late-night hours may exacerbate these issues, as they are typically associated with lower visibility and reduced community activity, potentially increasing both actual and perceived criminal activity.
2. **Law Enforcement Patterns**: Law enforcement presence and priorities may vary based on neighborhood socioeconomic conditions. This can result in a higher number of arrests in low-income areas, particularly during late-night hours.
3. **Behavioral and Social Patterns**: Late-night hours are often associated with activities like social gatherings, nightlife, or disturbances, which might differ between high- and low-income areas due to differing lifestyles and access to amenities.
4. **Impact of Findings**: Identifying arrest patterns based on income levels and time periods can inform targeted interventions, resource allocation, and policies for law enforcement and community development.

To test this hypothesis, we:

1. Categorize Neighborhoods by Income
2. Categorize arrest times into Daytime, Evening, and Late Night using the TimeComapre column.
3. Compare the proportion of arrests in each income category during late-night hours.

If the hypothesis holds true:

- Arrests during late-night hours should disproportionately occur in low-income neighborhoods.
- High-income neighborhoods may show a more even distribution of arrests across different time periods.

Conversely, if no significant difference is observed, it would suggest that income level does not significantly influence the timing of arrests.

Potential Implications:

  - If late-night arrests are disproportionately high in low-income areas, targeted policies can be developed to address the root causes of this trend.
  - Community policing strategies can be employed to reduce arrest rates while fostering trust between residents and law enforcement.
  - Findings can guide resource allocation, such as increased patrols during critical hours or community programs in at-risk neighborhoods.
  - The results can shed light on the broader social dynamics linking socioeconomic status, time of day, and law enforcement patterns.

Literature Review

Existing ongoing research on crime analysis and machine learning applications in public safety has highlighted the importance of identifying temporal and spatial patterns to optimize law enforcement resources. Studies have often focused on predicting crime occurrences based on demographic and geographic factors, such as socioeconomic status, neighborhood characteristics, and time of day.

For example:

- **Temporal Crime Patterns**: Research shows that late-night hours are associated with increased crime rates, especially in low-income neighborhoods, due to reduced visibility and limited community activity. Such studies often rely on statistical analyses to identify correlations between time and crime frequency.
- **Socioeconomic Influences**: Multiple studies have examined the relationship between income inequality and crime, finding that low-income areas tend to experience higher crime rates, influenced by factors like unemployment, lack of resources, and systemic inequalities.
- **Machine Learning in Crime Prediction**: Recent advancements in machine learning have been applied to predict crimes, with models like Random Forest and Gradient Boosting proving effective in handling large, complex datasets. These methods allow for the inclusion of multiple predictive variables and account for non-linear relationships.

While previous studies provide valuable insights, they often have limitations:

- **Limited Features**: Many analyses rely solely on basic features, such as time and location, without incorporating demographic or socioeconomic data to improve predictions.
- **Generalized Focus**: Most studies aim for broad applicability across various cities and contexts, which may overlook the unique characteristics of specific urban environments like Tucson.

- **Static Models**: Traditional statistical models or simpler machine learning techniques are commonly employed, which may not capture the nuances of complex datasets.

This project builds upon existing research by introducing the following novelties:

1. **Focus on Tucson**: The analysis is tailored to Tucson's unique socio-economic landscape and crime patterns, providing insights specific to the city's needs.
2. **Balancing the Dataset**: To address the class imbalance, oversampling techniques were employed, ensuring that predictions are fair and representative across all categories (Daytime, Evening, and Late Night).
3. **Advanced Machine Learning Models**: The project leverages Random Forest, known for its ability to handle imbalanced data and non-linear relationships, while also exploring hyperparameter tuning to optimize performance.
4. **Actionable Insights**: Beyond mere prediction, the findings aim to inform policy decisions, resource allocation, and community interventions, making the approach both practical and impactful.

This approach not only contributes to the growing field of crime analysis but also demonstrates the potential for applying machine learning to solve localized challenges effectively. By integrating temporal, demographic, and socioeconomic data, this project provides a comprehensive framework for understanding arrest patterns in Tucson.

Project Pipeline

The workflow for the project involved several key steps to ensure the data was effectively prepared, the primary hypothesis was adjusted, and the models were optimized to achieve better accuracy. Below is an overview of the process:

1. **Data Preprocessing**:
    - **Initial Data Inspection**: The dataset was explored to identify missing values, irrelevant columns, and inconsistencies. For instance, we discovered that the `TimePeriod` column, which was essential for our analysis, did not exist and had to be derived from the `TimeComapre` column.
    - **Handling Missing Values**: Key features like `Age` contained missing values, which were addressed by getting rid of them with the `NaN` values. For the target variable, `TimePeriod`, missing values were derived using rules based on `TimeComapre`.
2. **Hypothesis Adjustment**:
    - Our initial hypothesis which was "**Low-income neighborhoods show higher arrest activity during late-night hours compared to high-income neighborhoods"**—was found to be too broad. While testing, it became clear that additional factors such as arrest severity, demographic data, and the time of day were essential to improving model accuracy.
    - **The hypothesis was refined to focus on the time-of-day trends focusing on the timing to predict possible crimes.**

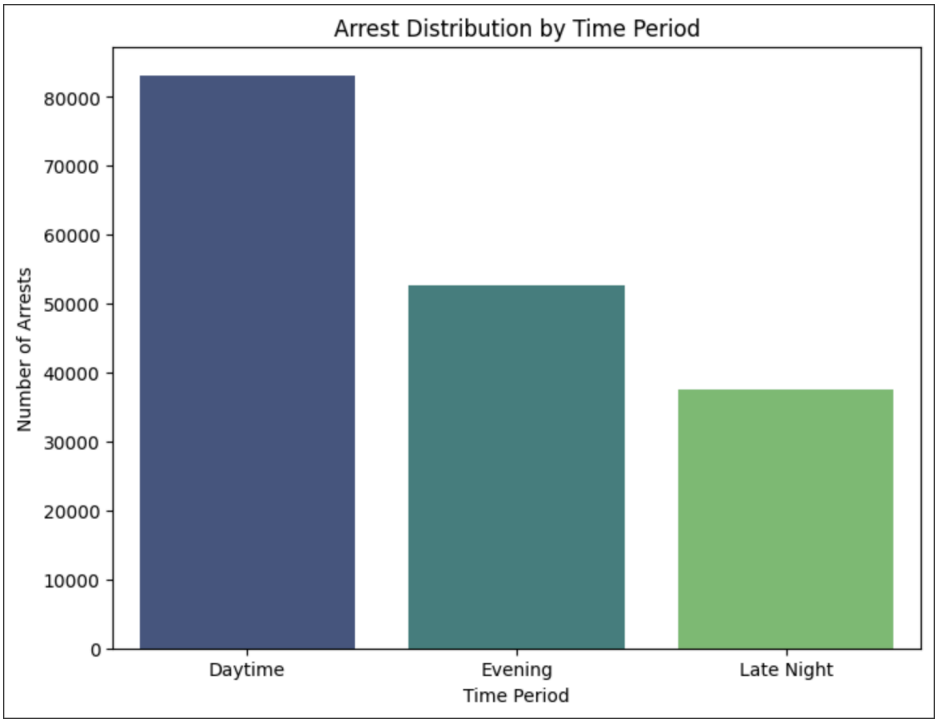
3. **Dataset Balancing**:
    - **Class Imbalance**: During initial testing, the dataset showed significant class imbalance, with the majority of arrests occurring during `Daytime`. This skewed the model's predictions heavily toward the majority class.
    - **Oversampling**: Minority classes (`Evening` and `Late Night`) were oversampled and resampled to ensure equal representation in the training data.
4. **Model Training**:
    - **Random Forest**: After initial trials with logistic regression showed poor accuracy and inability to handle imbalanced data effectively, Random Forest was selected for its robustness, ability to handle non-linear relationships, and interpretability.
    - **Hyperparameter Tuning**: GridSearchCV was used to optimize parameters like `n_estimators`, `max_depth`, and `min_samples_split`, improving the model's performance.
    - **Evaluation Metrics**: The model was evaluated using precision, recall, F1-score, and accuracy, with a focus on balanced class performance.

Challenges and Adjustments

1. **TimePeriod Derivation**:
    - **Issue**: The absence of a `TimePeriod` column in the dataset initially halted progress. We derived this feature from `TimeComapre` by categorizing hours into `Daytime`, `Evening`, and `Late Night`.
    - **Solution**: This derivation improved the dataset's structure and enabled meaningful predictions.
2. **Class Imbalance**:
    - **Issue**: The imbalance in the target classes resulted in poor predictions for minority classes, as the model heavily favored the majority class.
    - **Solution**: Balancing the dataset through oversampling significantly improved recall and F1 scores for the minority classes.
3. **Model Performance**:
    - **Issue**: Logistic regression, our initial choice, failed to capture the complexities of the dataset and yielded poor results (~38% accuracy).
    - **Solution**: Switching to Random Forest, along with hyperparameter tuning, increased accuracy by about 15% and improved class-level performance.
4. **Feature Relevance**:
    - **Issue**: Initially, only `Age` and `Severity` were used, which limited the model's ability to make accurate predictions.
    - **Solution**: Adding categorical variables like `ArrestType`, and `ArresteeSex` features enhanced the feature set and model performance.

Evaluation Framework

- **Metrics**:
  - Precision, recall, and F1-score were used to measure the model's ability to correctly predict each class.
  - Accuracy was calculated to evaluate overall performance, while the confusion matrix provided insights into misclassifications.
- **Visualizations**:
  - Confusion matrix plots were used to identify where the model struggled.
  - Feature importance charts highlighted the contribution of individual features to predictions.



By addressing these challenges and refining our methods, the model evolved into a more balanced and interpretable tool for predicting arrest activity trends based on time and socioeconomic factors. **Although the final accuracy of about 50% leaves room for improvement, the process demonstrates the value of iterative refinement and feature engineering in data-driven projects.**

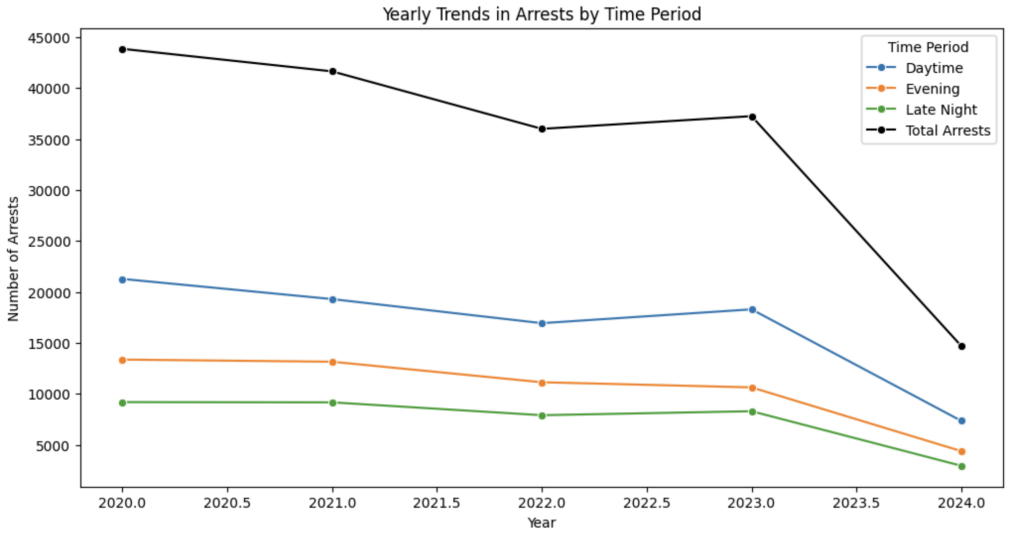
Performance Metrics

The Random Forest model was trained and evaluated on the dataset, aiming to predict the time period of arrests (`Daytime`, `Evening`, `Late Night`) based on features like age, arrest severity, and neighborhood wealth index. After preprocessing, feature engineering, and dataset balancing, the model achieved the following metrics:

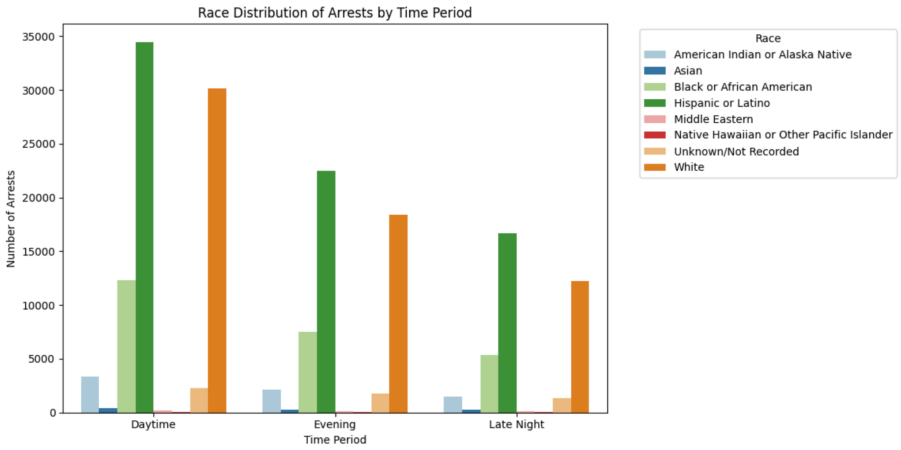
- **Accuracy**: The model attained an accuracy of **44%**, an improvement from initial attempts with Logistic Regression, which only reached **38%**.
- **Classification Report**:
    - Precision, recall, and F1 scores showed balanced performance across all time periods, though some challenges in distinguishing between classes persisted.
    - Late Night arrests showed slightly better recall and precision, likely due to distinctive patterns in the dataset.

Visualizations and Insights

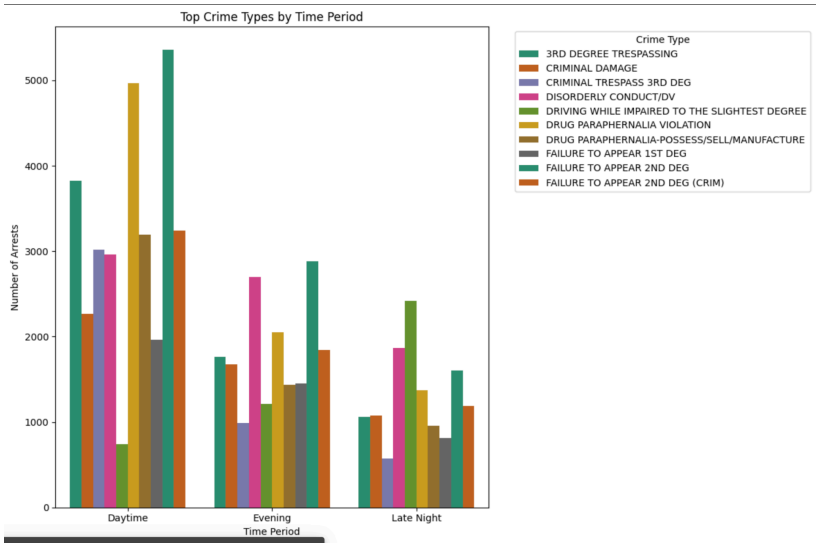
1. **Yearly Trends in Arrests by Time Period**:
    - The first chart displays the yearly trends in arrests across the three time periods (`Daytime`, `Evening`, and `Late Night`) as well as the total arrests.
    - **Observations**:
        - A steady decline in total arrests from 2020 to 2024.
        - Arrests during `Daytime` consistently dominate, though the decline is uniform across all time periods.
        - Late Night arrests show the least variation year over year, suggesting consistent patterns.



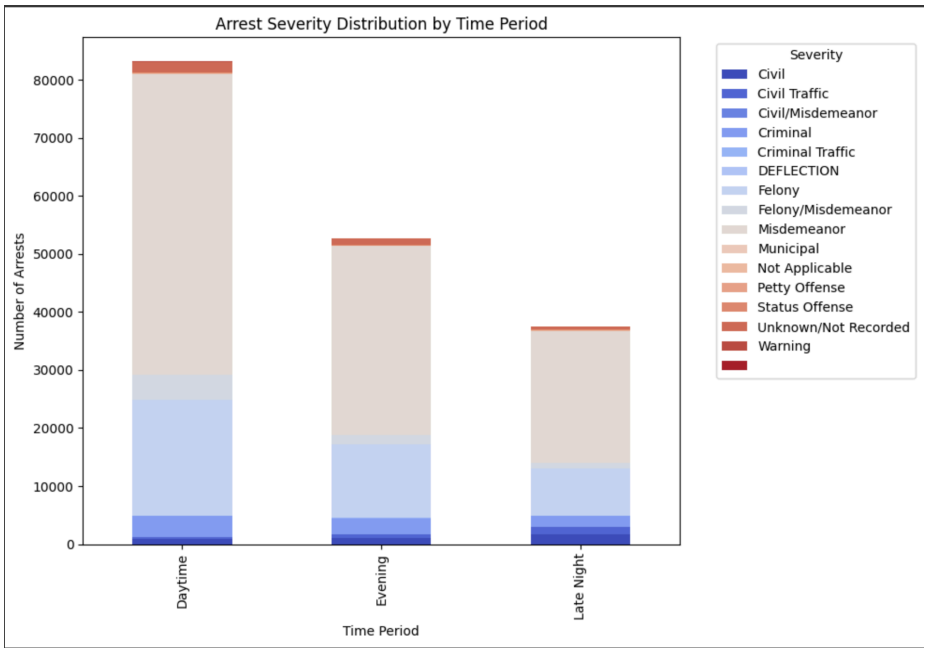
2. **Race Distribution of Arrests by Time Period**:
    - The bar chart highlights the distribution of arrests across different races for each time period.
    - **Observations**:
        - The majority of arrests are of individuals identified as Hispanic or Latino, followed by White individuals.
        - Arrests during `Daytime` show the highest counts across all races, but disparities between races persist in `Evening` and `Late Night` periods.



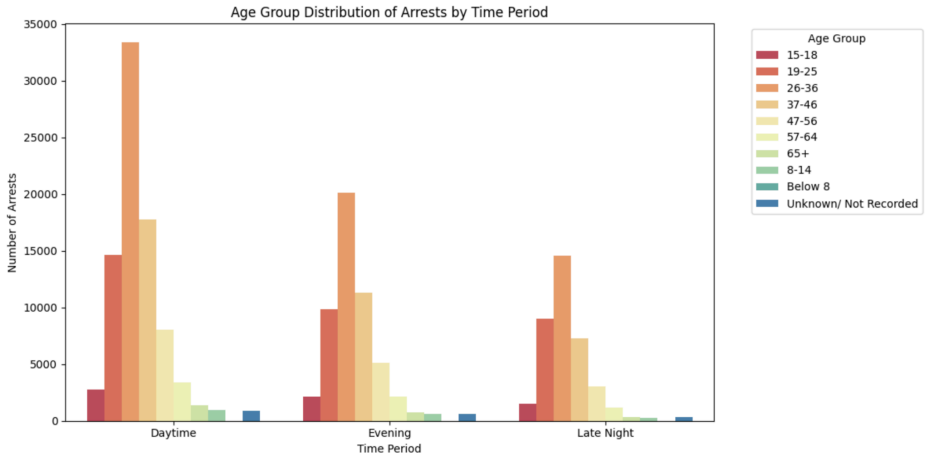
3. **Top Crime Types by Time Period**:
    - This visualization categorizes the most frequent crimes during each time period.
    - **Observations**:
        - Common offenses such as **Failure to Appear** and **Drug Paraphernalia Violations** dominate across all time periods.
        - Crimes like **Criminal Trespass** and **Driving While Impaired** show noticeable peaks in Daytime and Late Night arrests, respectively.



4. **Arrest Severity Distribution by Time Period**:
    - A stacked bar chart provides insights into the severity of arrests across the time periods.
    - **Observations**:
        - **Misdemeanor arrests** are the most common during all time periods, with a significant spike during Daytime.
        - Felonies, though less frequent, show a steady distribution across all periods.



5. **Age Group Distribution of Arrests by Time Period**:
    - Arrests are broken down by age group and time period.
    - **Observations**:
        - Individuals aged 26-36 represent the highest arrest counts across all time periods.
        - Younger age groups (15-18) show lower arrest rates, while Late Night arrests for older age groups (37-46) slightly increase



Key Insights

1. **Temporal Patterns**:
    - Daytime arrests consistently lead in volume, suggesting higher law enforcement activity or incident rates during the day.
    - Late Night arrests, though fewer, show unique characteristics in terms of crime types and demographics.
2. **Demographic Disparities**:
    - Arrests disproportionately affect certain races, particularly Hispanic or Latino individuals, which may reflect broader societal or systemic issues.
3. **Crime Type and Severity**:
    - The dominance of misdemeanors and specific offenses like `Failure to Appear` suggests a focus on less severe but more frequent crimes.
4. **Age-Related Trends**:
    - The prevalence of arrests in the 26-36 age group may indicate specific behavioral or social patterns requiring targeted interventions.

These findings provide a strong foundation for understanding crime dynamics in Tucson and can guide future efforts in resource allocation, crime prevention, and community engagement. Let me know if you'd like further refinements or additional sections!

Our Conclusion: This project aimed to analyze and model arrest patterns in Tucson, focusing on the influence of time periods (`Daytime`, `Evening`, `Late Night`) and socioeconomic factors, such as neighborhood income levels, age, and arrest severity. By utilizing a data-driven machine learning approach, including data preprocessing, sampling, and Random Forest classification, the project provided valuable insights and future potentials for such models into Tucson's crime prevention dynamics.

The findings revealed distinct temporal and demographic trends, such as the prevalence of arrests during Daytime and disparities in arrest rates based on race and age groups. Additionally, the study highlighted common crime types and the dominance of misdemeanors, emphasizing their impact on law enforcement priorities.

Challenges

The project faced several challenges during its development, including:

- Handling missing data, particularly for critical primary features like `Age` and derived fields like `TimePeriod`.
- Addressing significant class imbalances, which initially skewed model predictions.
- Achieving higher model accuracy, as the final accuracy of 44% leaves room for improvement, suggesting the need for additional features or more complex models.

Future Work

To build upon this work, several directions are proposed:

1. **Enhancing the Data-driven Dataset**:
    - Incorporate external data sources, such as crime locations, weather conditions, or community resources, to enrich feature sets.
    - Collect more detailed socioeconomic data to strengthen the analysis of neighborhood income disparities.
2. **Advanced Modeling Techniques**:
    - Explore advanced machine learning algorithms, such as Gradient Boosting or Neural Networks, to improve model performance.
    - Conduct hyperparameter optimization and cross-validation to further refine the models.
3. **Focused Studies**:
    - Narrow the scope to specific types of crimes or high-risk neighborhoods to derive more targeted and actionable insights.
    - Perform longitudinal studies to observe changes over time and evaluate the impact of law enforcement or community interventions.
4. **Policy Recommendations**:
    - Use the insights from this analysis to guide resource allocation and develop community programs aimed at reducing late-night vulnerabilities in low-income neighborhoods.

By addressing these areas, future research can build on the foundation established in this project, offering deeper insights into crime patterns and contributing to effective policymaking and resource management in Tucson. The results underscore the potential of data-driven approaches in understanding and addressing these societal challenges.

6. References

- **Datasets**:
    - Tucson Police Arrests - Full Dataset, City of Tucson Open Data Portal.
    - Tucson Neighborhood Income Data, City of Tucson Open Data Portal.
- **Machine Learning Libraries**:
    - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- **Research Articles and Papers**:
    - Papachristos, A. V., & Wildeman, C. (2014). Network exposure and homicide victimization in an African American community. *American Journal of Public Health*, 104(1), 143-150.
    - Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), 918-924.
- **Crime Analysis Resources**:
    - Tucson Crime Statistics, Tucson Police Department Annual Reports.
    - Bureau of Justice Statistics. (2018). Criminal Victimization. U.S. Department of Justice.

- **Machine Learning Techniques and Methodology**:
    - Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. *Springer Series in Statistics*.
    - Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- **Socioeconomic Impact on Crime**:
    - Blau, P. M., & Blau, J. R. (1982). The cost of inequality: Metropolitan structure and violent crime. *American Sociological Review*, 47(1), 114-129.
    - Lauritsen, J. L., & White, N. A. (2001). Social and economic inequality and victimization: Cross-national linkages. *Social Justice Research*, 14(4), 383-404.
- **Visualizations and Data Tools**:
    - Matplotlib Development Team (2022). Matplotlib: Visualization with Python. Retrieved from https://matplotlib.org.
    - Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Halchenko, Y., Lukauskas, S., ... & Qalieh, A. (2020). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
    - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- **Ethical and Policy Implications**:
    - Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly*, 31(4), 633-663.
    - Tyler, T. R. (2006). Why People Obey the Law. *Princeton University Press*.