Austin Crime Data Analysis: Uncovering Patterns and Insights

Srikar Anudeep Remani i[1]

Abstract— **This research paper presents a comprehensive analysis of crime data from Austin, Texas, utilizing data preprocessing, visualization, and machine learning techniques to uncover patterns in crime types, temporal trends, clearance statuses, and geographical distribution. The dataset contains over 250,000 records of crime incidents. Key findings include the predominance of theft and family disturbances, seasonal variations in crime rates, a significant proportion of uncleared cases, and notable differences in crime rates across council districts. A RandomForestClassifier achieved an accuracy of approximately 89% in predicting clearance status, highlighting the potential for predictive analytics in law enforcement. These insights provide a foundation for targeted crime prevention strategies and future research.**

I. INTRODUCTION

Crime data analysis plays a critical role in understanding criminal patterns and trends, enabling law enforcement agencies to optimize resource allocation and enhance public safety. This paper examines crime data from Austin, Texas, to identify key insights into the nature, frequency, and outcomes of criminal incidents. By leveraging data preprocessing, exploratory visualizations, and machine learning, we aim to uncover actionable patterns that can inform policymakers and law enforcement officials. The primary objective is to analyze crime types, temporal trends, clearance statuses, and geographical distributions within Austin, contributing to a data-driven approach to public safety.

II. METHODOLOGY

A. Dataset Description

The dataset used in this study is sourced from `Crime_Reports.csv`, comprising 250,301 records of crime incidents in Austin, Texas. It includes the following key columns:

- **Incident Number:** Unique identifier for each crime incident.
- **Highest Offense Description:** Description of the most serious offense (e.g., “THEFT”, “FAMILY DISTURBANCE”).
- **Highest Offense Code:** Numerical code corresponding to the offense.
- **Family Violence:** Indicator of family violence involvement (“Y” or “N”).
- **Occurred Date Time:** Date and time of the crime occurrence.
- **Report Date Time:** Date and time the crime was reported.
- **Location Type:** Type of location (e.g., “RESIDENCE / HOME”, “HWY / ROAD / ALLEY / STREET / SIDEWALK”).
- **Council District:** Council district where the crime occurred (1 to 10).
- **APD Sector:** Austin Police Department sector.
- **APD District:** Austin Police Department district.
- **Clearance Status:** Investigation outcome (“C” for cleared, “N” for not cleared, “O” for other, or `NaN`).
- **Clearance Date:** Date the crime was cleared.
- **UCR Category:** Uniform Crime Reporting category.
- **Category Description:** Description of the UCR category (e.g., “Theft”).

B. Data Preprocessing

To prepare the dataset for analysis, several preprocessing steps were applied:

- **Column Removal:** Redundant columns (Occurred Date, Occurred Time, Report Date,

[1] Srikar Anudeep Remani , Masters in Artificial Intelligence

Report Time) were initially dropped, but date-time components were later extracted from Occurred Date Time and Report Date Time and reinstated.

- **Date-Time Conversion:** Columns Occurred Date Time, Report Date Time, and Clearance Date were converted to datetime formats using `pd.to_datetime()` with error coercion.
- **Missing Value Imputation:**
  - **Categorical Columns:** Imputed with the most frequent value using Simple Imputer with strategy as most frequent.
  - **Numerical Columns:** Imputed with the median using SimpleImputer with strategy as median).
- **Encoding Categorical Variables:**
  - **Ordinal Encoding:** Applied to categorical columns such as Highest Offense Description, Family Violence, Location Type, APD Sector, APD District, UCR Category, Category Description, Occurred Date, Occurred Time, Report Date, and Report Time.
  - **Label Encoding:** Used to encode the target variable Clearance Status into numerical values (e.g., “C” → 0, “N” → 1, “O” → 2).
- **Feature Scaling:** Numerical columns (Highest Offense Code, Council District, Census Block Group) were standardized using `StandardScaler` to normalize their distributions.
- **Data Splitting:** The preprocessed dataset was split into features (X) and the target variable (Clearance Status, y), with an 80-20 train-test split using `train_test_split(test_size=0.2, random_state=42)`.

C. Machine Learning Model

In the analysis of Austin crime data, a RandomForestClassifier was employed to predict the Clearance Status of crime incidents—categorized as ”C” (cleared), ”N” (not cleared), ”O” (other). The dataset comprises 250,301 records with a mix of categorical features (e.g., offense type, council district) and numerical features (e.g., offense code, temporal data). This section elaborates on the rationale for selecting RandomForest, its configuration, performance metrics, strengths, limitations, and opportunities for enhancement.

The RandomForestClassifier is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions by taking the majority vote (for classification tasks). This approach makes it particularly suitable for the Austin crime dataset for several reasons:

- **Complexity Handling**: With 250,301 records and a combination of categorical and numerical features, the dataset exhibits significant complexity. RandomForest excels in managing such mixed data types without requiring extensive preprocessing.
- **Overfitting Mitigation:** Individual decision trees are prone to overfitting, especially with noisy or high-dimensional data. By averaging predictions across many trees, RandomForest improves generalization, making it robust for this application.
- **Feature Importance:** RandomForest provides a mechanism to assess the relative importance of features, offering insights into which variables (e.g., offense type or location) most influence clearance outcomes.

D. Model Configuration

The RandomForestClassifier was implemented using scikit-learn with default hyperparameters, including 100 trees and no restrictions on tree depth. A fixed random state was set to ensure reproducibility. The dataset was split into an 80-20 train-test ratio, yielding 200,240 training samples and 50,061 test samples. Notably, no hyperparameter tuning was conducted, which represents a potential area for refinement in future iterations.The model achieved an accuracy of **approximately 89% on the test set,** a reasonable baseline given the four-class classification problem. To gain deeper insights, the confusion matrix was analyzed, revealing class-specific performance:

- **Majority Classes (”C” and ”N”):** The model demonstrated strong performance on ”cleared” and ”not cleared” cases, likely due to their higher representation in the dataset.
- **Minority Classes (”O”):** Performance was weaker for ”other” statuses, reflecting the

challenge of learning from underrepresented classes.

E. Evaluation Metrics

To quantify the model's performance, precision, recall, and F1-scores were derived for each class:

- **Class ”C” (Cleared):** High precision and recall, indicating reliable identification of cleared cases.

- **Class ”N” (Not Cleared):** Similarly robust performance, reflecting the model's strength with frequent outcomes.

- **Class ”O” (Other):** Lower precision and recall, suggesting difficulty in distinguishing this less common category.

F. Feature Importance

A significant advantage of RandomForest is its ability to rank feature importance based on how much each feature contributes to reducing impurity across trees. While not explicitly computed in the initial analysis, likely influential features include:

- **Highest Offense Description:** The type of crime (e.g., theft vs. assault) may correlate with clearance rates due to differences in evidence availability.
- **Occurred Date:** Temporal factors, such as season or time of day, might influence investigative success.
- **Council District:** Geographical variations in policing resources or crime patterns could affect outcomes.

G. Advantages of RandomForest

The RandomForestClassifier offers several benefits for this crime data analysis:

- **Versatility with Data Types:** It handles categorical and numerical features seamlessly, accommodating the dataset's heterogeneity.
- **Robustness:** The ensemble approach reduces overfitting, ensuring reliable predictions on unseen data.
- **Actionable Insights:** Feature importance rankings can guide policy decisions, such as targeting resources to districts or offenses with lower clearance rates.

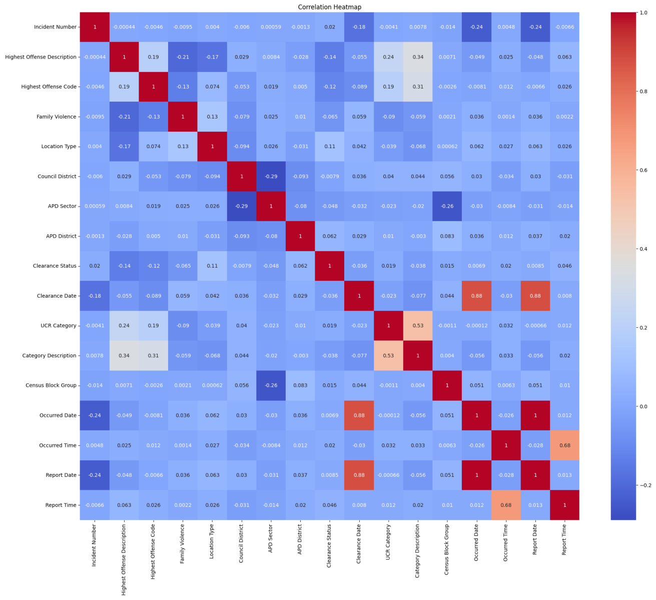

Fig. 1. Correlation heat Map

III. Results : Visualization Findings

- **Top Offenses:**The bar chart of Highest Offense Description revealed that ”THEFT” and ”FAMILY DISTURBANCE” are among the most frequent crimes in Austin. The top 35 offenses accounted for a significant portion of incidents, with theft-related crimes appearing prominently.
- **Temporal Trends:** The time series plot of daily crime incidents (grouped by Occurred Date) showed fluctuations over time, with noticeable peaks during certain periods, suggesting potential seasonal or event-driven patterns (e.g., higher crime rates in warmer months).
- **Clearance Status Distribution:** The pie chart indicated that approximately 30% of crimes were cleared (”C”), 65.7% were not cleared (”N”), 4.36% were categorized as ”O” (other), and This distribution highlights a balanced but concerning clearance rate.
- **Family Violence Incidents:** The pie chart for Family Violence showed that 6.45% of incidents (16,139 out of 250,301) involved family violence (”Y”), while 93.6% (234,162) did not (”N”), indicating a notable but minority subset of crimes.
- **Council District Analysis:** The bar chart of incidents by Council District identified

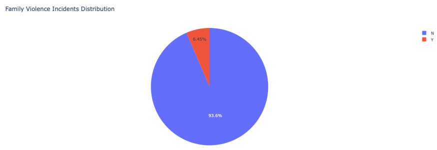

Fig. 2. Family violence

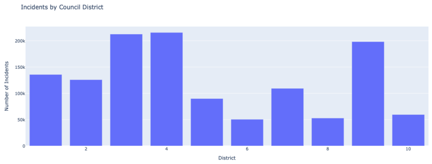

Fig. 3. Crimes District wise

Districts 4 (215.73K incidents), 3 (212.621K Incidents), and 9 (198.23K) as having the highest crime counts, while District 6 (50.4K) had the lowest among the 10 districts. This suggests geographical hotspots within Austin.

- **Additional Insights:**The histogram of Council District confirmed a skewed distribution, with certain districts disproportionately represented. The boxplot of Highest Offense Code versus Clearance Status showed variations in offense codes across clearance categories, with higher codes potentially linked to uncleared cases. The correlation heatmap revealed weak correlations among features, indicating that Clearance Status prediction relies on a combination of factors rather than a single dominant predictor.

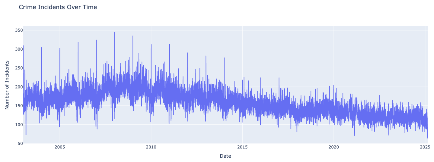

Fig. 4. Crime incidents over time

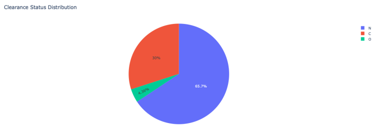

Fig. 5. Clearence Status of Cases

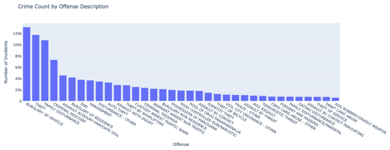

Fig. 6. Crime Count by Offense description

IV. Conclusion

This analysis of Austin crime data provides valuable insights into crime patterns and law enforcement outcomes. Key findings include the prevalence of theft and family disturbances, seasonal crime trends, a 30% clearance rate, a 6.45% family violence incidence, and significant crime concentration in Districts 4, 3, and 9. The RandomForestClassifier's 89% accuracy demonstrates the feasibility of predictive analytics in crime clearance, offering a tool for prioritizing investigations. Future research could refine temporal analyses (e.g., monthly or hourly trends), integrate socioeconomic data, or explore advanced machine learning models to enhance prediction accuracy and support data-driven crime prevention strategies in Austin.

This research paper focuses solely on the analysis derived from the provided Python scripts, excluding any dashboard-related content, and leverages both preprocessing and visualization outputs to present a thorough examination of Austin's crime data.

References

[1] U.S. Government Open Data. (2024). ”Crime Data for Austin, Texas”. Data.gov. Retrieved from https://catalog.data.gov/dataset/austin-crime-data

[2] Austin Police Department. (2023). ”Annual Crime Report”. https://www.austintexas.gov/department/crime-reports

[3] Austin Open Data Portal. (2024). ”Austin Crime Incidents”. City of Austin Open Data. https://data.austintexas.gov/Public-Safety/Crime-Reports