VISUAL ANALYTICS AND BUSINESS ANALYTICS - BAN5573

Crime Prediction with Data-Driven Intelligence

Professor: Hamidreza Ahady Dolatsara

Submission Date: 04/08/2025

Group Members:

Ritu Adhikari

Pragathi Somashekaraiah

Ariff Khan

Rishabh Prasad

Motivation

Crime remains a critical problem in increasingly complex and dynamic urban environments with social, economic, and psychological implications. While traditional crime prevention methods remain important, on their own they are no longer sufficient. Instead, data-driven methods, enhanced by improvements in statistical and machine learning methods, will serve as a powerful addition to traditional policing methods. This project will provide the analytical tools to examine underlying patterns in crime data, as well as help support evidence-based processes in choosing options. The premise for this study is that public safety remains a growing priority, and many municipalities want better use of policing resources. After identifying communities at risk and forecasting crime, municipalities can implement targeted interventions, better deploy resources, and prevent future crime. The project will contribute to addressing important questions such as: What are predictors of crime? Can we reliably forecast it?

Introducing the Topic

Crime analysis has come to the forefront as urban populations expand. This project, Crime Prediction with Data Driven Intelligence, analyzes crime data throughout the country (2018-2024), applying descriptive and predictive analytics to explore patterns, trends, and anomalies in crime. Through a combined use of statistical techniques, geospatial mapping, and predictive modeling within Tableau, this project will provide a unique, interactive view of crime in the U.S.

Project Scope

Data Collection & Preprocessing: Collecting incident-level records from open-source data, normalizing fields (such as date and region), managing missing values, and deriving temporal features (e.g., month, year, and potentially day of week).

Descriptive Analysis: Reviewing key crime metrics through summary statistics and visualizations. Exploring frequency across geographic locations and temporal dimensions by crime type to uncover trends and hotspots.

Predictive Modeling: Creating and evaluating machine learning models for crime prediction, including classification (to identify crime types) and regression (to predict the number of incidents). Performance will be assessed using cross-validation, accuracy scores, RMSE, and feature importance.

Recommendation System: Translating model outputs into risk-based strategies. Recommending patrol schedules, optimized resource allocation for high-risk zones, and community outreach programs based on crime forecasts.

Reporting & Visualizations: Delivering the final analysis through dashboards and reports. Using heatmaps, time series charts, and geographic region-based maps to communicate insights clearly to stakeholders.

Data Source and Description

Data Source: The dataset used in this project is sourced from the Realtime Crime Index, a comprehensive open-data platform aggregating incident-level reports from law enforcement agencies across the United States. It includes over 40 million records with details on crime categories, location (state, region), and population coverage.

Description of the Dataset

- Month: The month when the crime data was recorded.
- Year: The year of the data entry.
- Date: Likely a full timestamp (e.g., 2024-02-01), combining month and year.
- State: U.S. state abbreviation (e.g., NY, CA).
- Region: U.S. census region (e.g., Northeast, Midwest, South, West).
- Murder: Number of murders reported.
- Rape: Number of rapes reported.
- Robbery: Number of robberies.
- Aggravated Assault: Cases of aggravated assault.
- Burglary: Incidents of unlawful entry.

- Theft: Non-violent thefts like shoplifting or larceny.
- Motor Vehicle Theft: Stolen vehicles.
- FBI.Population.Covered: Number of people covered by the data collection for that agency.
- Number.of.Agencies: Count of agencies contributing to the report in that jurisdiction.
- Total_Incidents: Total number of reported incidents for all crime types combined in that record (summation of individual crime types).
- Crime_Cluster: A derived or modeled field —from a clustering algorithm (like K-Means) that categorizes regions into crime levels/types (e.g., high-crime, moderate, low-crime areas)

Literature Review

In recent years, crime analysis has shifted toward a focus on data-driven intelligence, with predictive analytics and machine learning at the center of current prediction and forecasting systems used to predict policing trouble areas. Crime prediction is not limited to hot spots; it also includes classification and regression of crime-type, based on historical patterns.

Spatio-temporal modeling has become a central theme in intelligent crime analytics. Research by Mohler et al. (2014) introduced self-exciting point processes (SEPP) that modeled crime occurrence similarly to earthquake aftershocks, suggesting that past crimes significantly influence future crime probability. These models demonstrated notable success in short-term hotspot prediction.

Machine learning approaches, including decision trees, support vector machines, and deep learning, have been tested for classification tasks such as predicting crime categories based on contextual features (Wang & Brown, 2019). Logistic regression and random forest models have also been widely adopted for crime rate forecasting due to their high interpretability and accuracy. For instance, Huang et al. (2021) showed that crime rates could be effectively predicted using ensemble learning when trained on location, time, weather, and socio-demographic data.

In addition to model building, visualization and geospatial tools have played a vital role in presenting findings to stakeholders. GIS-integrated dashboards allow law enforcement agencies to interactively monitor crime trends and optimize patrol allocations (Liu et al., 2020). The integration of real-time data into such platforms enhances operational response time and enables dynamic policing strategies.

While much of the literature supports predictive policing, several scholars caution against over-reliance on algorithmic output without considering underlying biases in historical data. Lum & Isaac (2016) argue that historical policing practices can skew prediction outcomes, especially in communities disproportionately affected by surveillance.

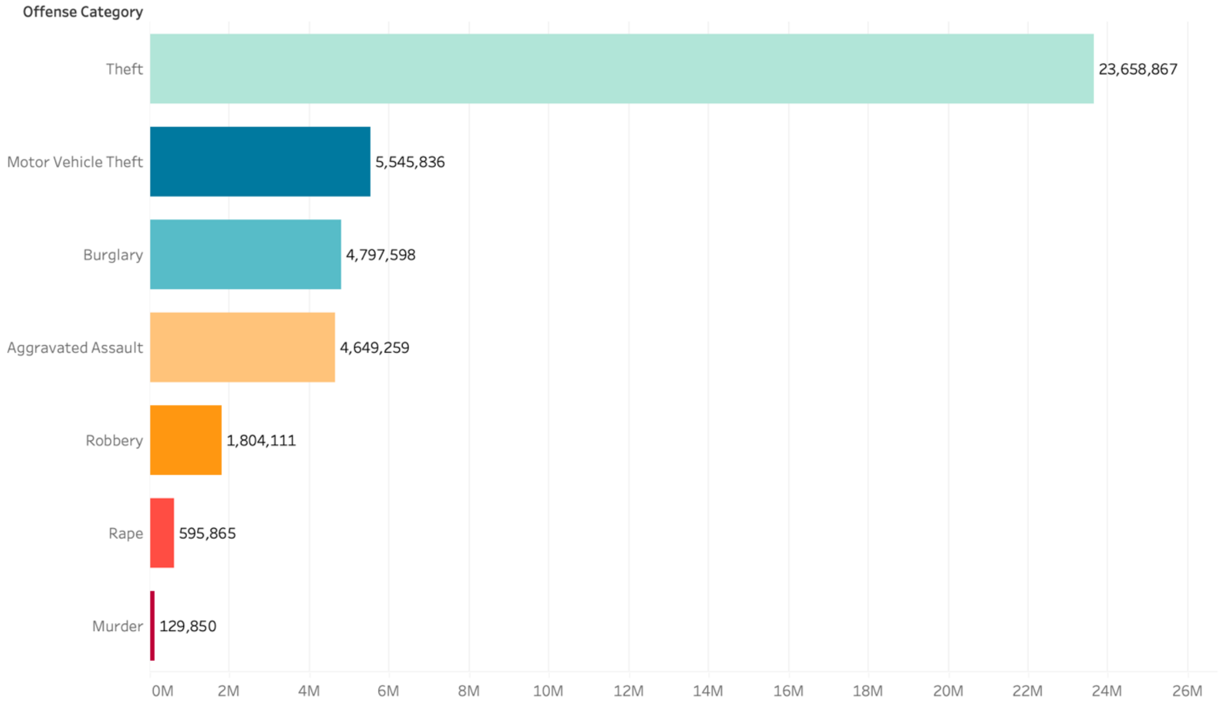
Overall, the literature supports the strategic integration of machine learning, spatio-temporal analysis, and interactive dashboards in modern crime analysis. These methodologies not only help in identifying crime patterns but also contribute to resource planning, proactive law enforcement, and policy formation—objectives closely aligned with the goals of this project.

Significance of the Study

Importance of the StudyThis study facilitates data-driven crime prevention through pattern identification and predictive analytics. This information allows law enforcement to place their resources in an efficient and effective manner, and enhances the safety of our citizens, and informs urban policy planning. Further, the information can instigate forward-thinking strategies, reduce crime, and leverage the communities to become smarter and safer.

Visualization Using Tableau

1- Types of Crime Distribution



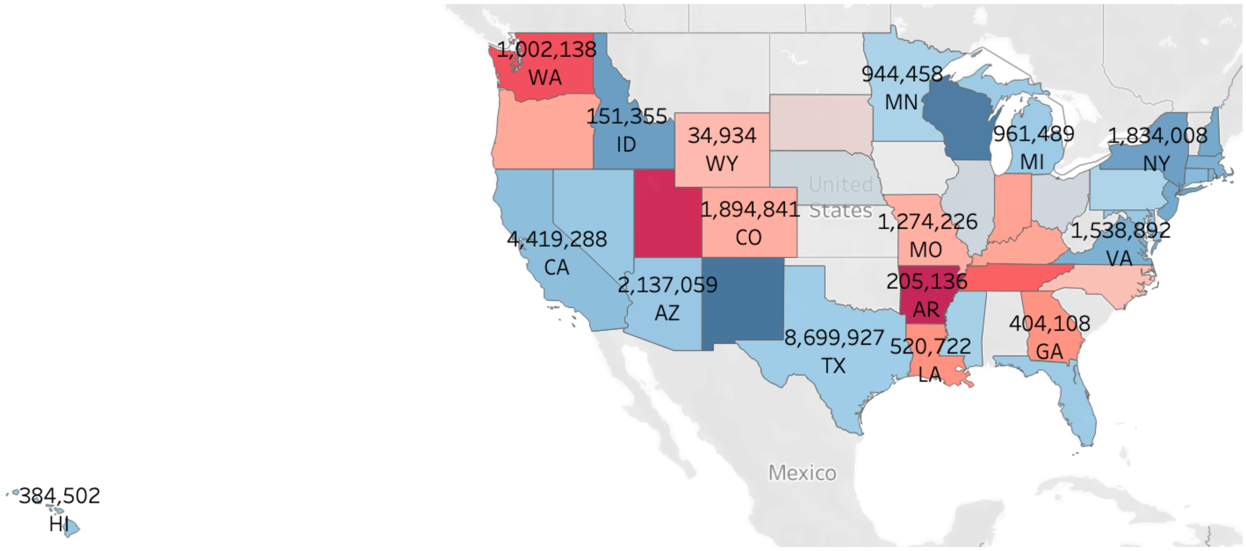
Explanation

This visualization shows the distribution of crime types across the dataset. Theft is by far the most prevalent crime, with over 23 million incidents, followed by Motor Vehicle Theft, Burglary, and Aggravated Assault — each ranging between 4 to 5.5 million cases. Violent crimes like Robbery, Rape, and Murder, while serious in nature, account for a significantly smaller proportion. This indicates a national crime pattern where property crimes dominate over violent offenses.

Real-Time Usage

Understanding which crimes occur most frequently can help law enforcement and city planners prioritize resource allocation. For example, the high frequency of theft-related incidents may prompt investment in public awareness campaigns, surveillance systems, or targeted patrolling in theft-prone areas. Simultaneously, even though Murder and Rape have lower incident counts, their severity demands continued focus in policy, prevention, and victim support services, especially in high-crime zones.

2- Crime by State



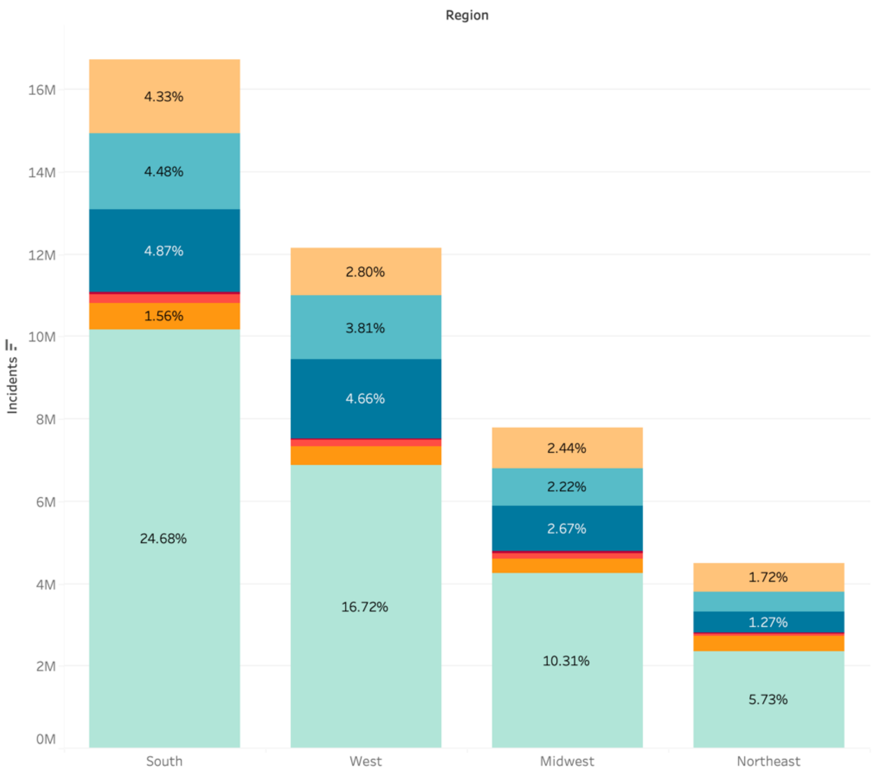
Explanation

This choropleth map displays the total number of reported crime incidents across U.S. states from 2018 to 2024. Texas leads with over 8.6 million incidents, followed by California and Arizona. In contrast, states like Wyoming and Hawaii have relatively low crime counts. The color gradient highlights crime intensity geographically, with darker shades indicating higher crime volumes. This spatial distribution offers a macro-level understanding of regional crime disparities, reflecting variations in population density, urbanization, and policing infrastructure.

Real-Time Usage

State-wise crime visualization helps policymakers and law enforcement agencies tailor region-specific strategies. For high-crime states like Texas and California, resource allocation, law enforcement staffing, and community programs can be prioritized. Additionally, this insight allows cross-state comparisons, prompting data-backed collaboration among jurisdictions. Urban planners and local governments can also utilize this data to integrate safety considerations into city development, public transport planning, and emergency response infrastructure.

3- Crime by Region & Type



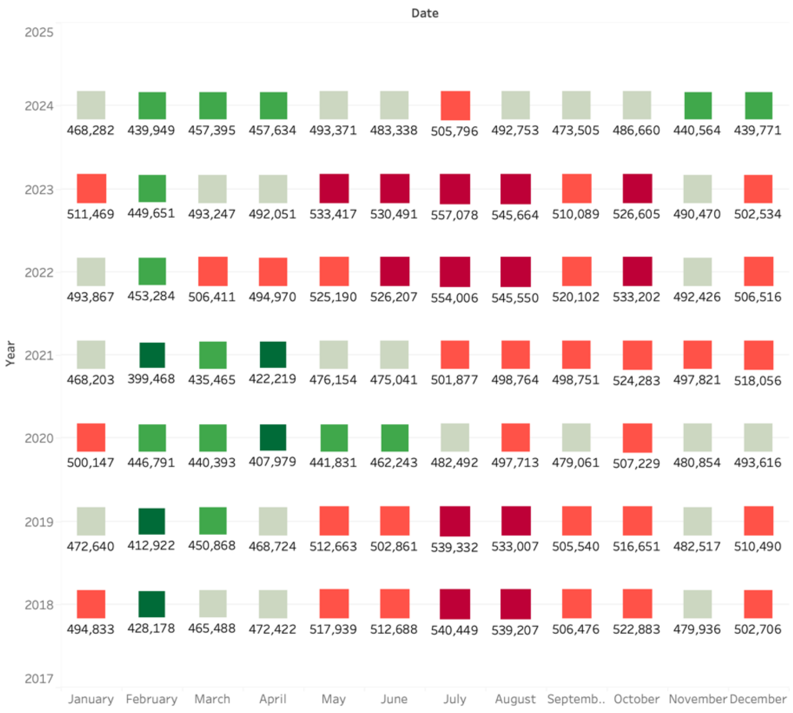
Explanation

This stacked bar chart shows how incidents of crime are spread across four U.S. regions, namely South, West, Midwest, and Northeast, broken out by category of offense. Based on the total number of incidents, Southern region has the most incidents of crime, particularly Theft (24.68%), followed by Western region. Overall, the Northeast has the fewest incidents of crime, but still a measurable percentage in the volume of Burglary and Aggravated Assault.

Real-time Usage

This breakdown can serve law enforcement agencies and city planners in their decision-making regarding the allocation of patrol resources and personnel deployment, according to the offense profile in the region. For example, the South might need to emphasize targeted theft-prevention tactics while the West may not need to. This visual aid will also help identify crime strategies at the region level and help inform targeted public safety marketing campaigns**.**

4- Monthly Seasonality Trends



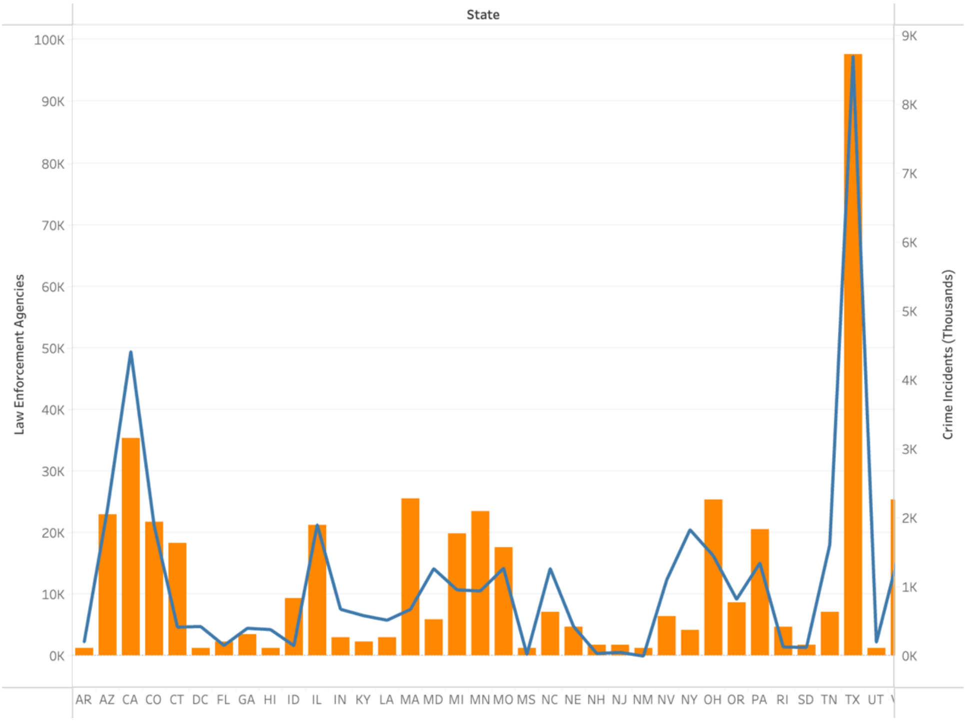
Explanation

The heatmap presents patterns in crime volume by month from 2018 to 2025. Each square represents the total amount of incidents that were reported during a month in a year, while coloring illustrates the relative volume of crime, where darker colors indicate more volume. The heatmap identifies a consistent seasonal pattern, with July and August displaying higher volumes of crime, which may indicate a tendency for crime to occur in warmer months. Incidentally, months such as February and April demonstrate lower volumes of crime than other months, representing the safest times of the year.

Real-time Usage

This trend analysis supports seasonal forecasting and preventive planning. Law enforcement can increase patrol deployment during peak months such as July and August and reduce operational intensity during calmer periods. Policy makers and city planners can also use these insights to schedule public safety campaigns, community outreach programs, or resource reallocations in anticipation of monthly crime surges, making urban areas safer and resource usage more efficient.

5- Agencies vs Crime



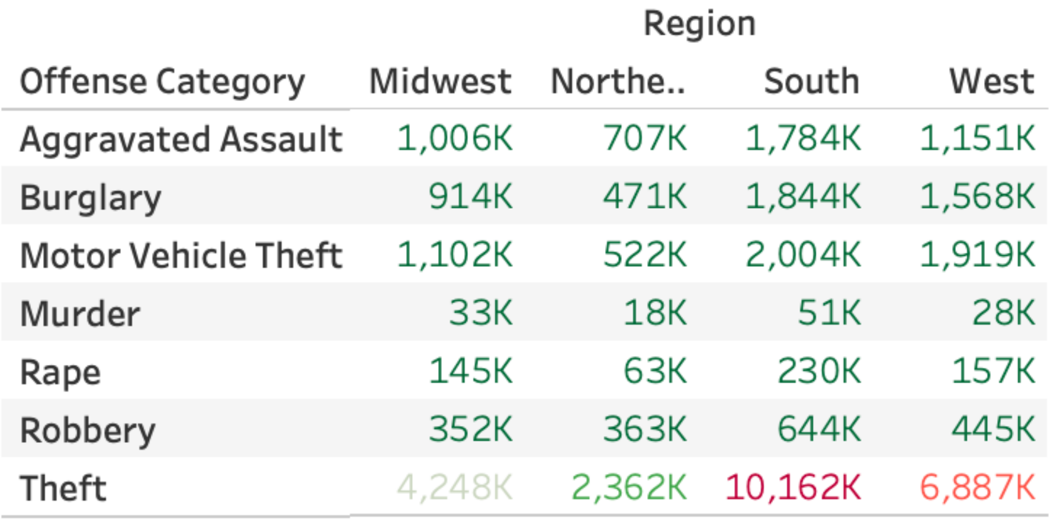
Explanation

This dual-axis chart compares law enforcement agency count (bar chart) with crime incidents in thousands (line chart) across U.S. states. It reveals major imbalances—for example, Texas has the highest crime incidents, while some states like California and Illinois show high agency counts with moderate incident levels. States like Wyoming or Vermont reflect both low agency presence and low incident volume. The visual emphasizes how crime rates and agency coverage do not always scale together, suggesting inefficiencies or regional disparities in resource deployment.

Real-time Usage

This visualization can guide policy and resource allocation by identifying where law enforcement might be over- or under-resourced relative to crime levels. Cities or states with high crime but fewer agencies might benefit from increased funding, recruitment, or tech-driven support. It also helps planners evaluate response readiness, staffing needs, and coverage density, enabling smarter deployment strategies for more effective crime prevention and public safety enhancement.

6- Resource Allocation Simulation



	Region			
Offense Category	Midwest	Northe..	South	West
Aggravated Assault	1,006K	707K	1,784K	1,151K
Burglary	914K	471K	1,844K	1,568K
Motor Vehicle Theft	1,102K	522K	2,004K	1,919K
Murder	33K	18K	51K	28K
Rape	145K	63K	230K	157K
Robbery	352K	363K	644K	445K
Theft	4,248K	2,362K	10,162K	6,887K

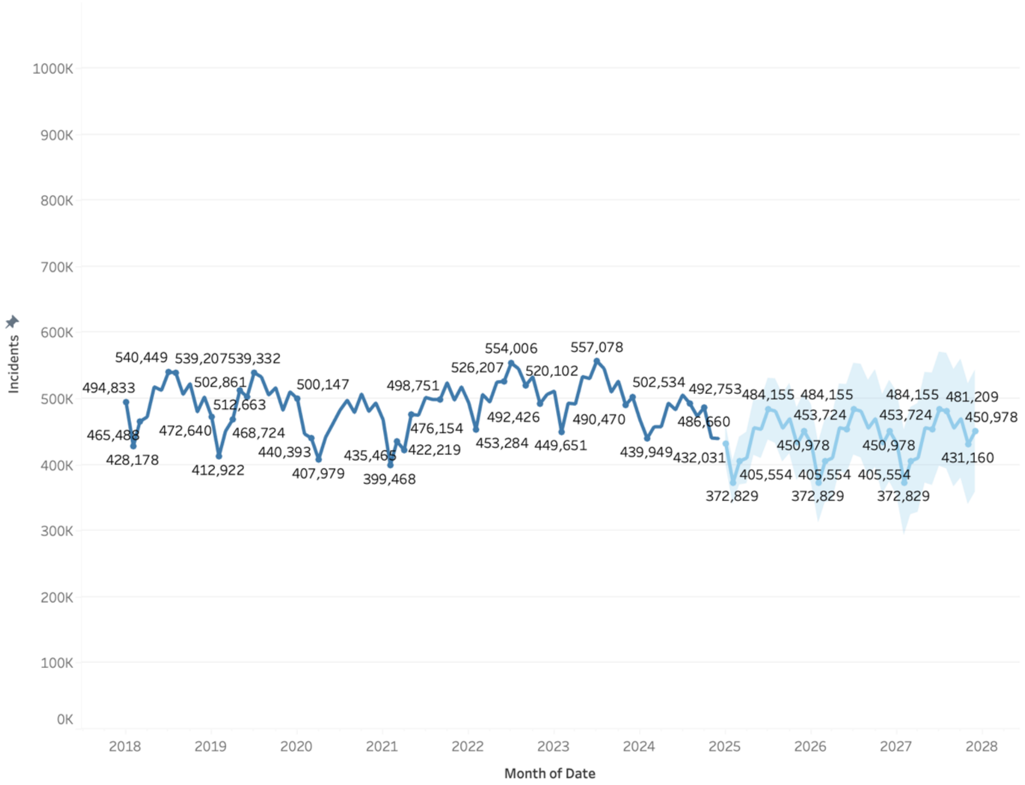
Explanation

This simulation table projects crime incidents by offense category across U.S. regions, adjusting for changes in law enforcement agency counts using a dynamic parameter (% increase). Color shading highlights the relative intensity of simulated crime, with darker shades signaling higher predicted incidents. For example, even after increasing agency presence, the South continues to show a large simulated burden, particularly in Theft and Motor Vehicle Theft, reflecting deeply entrenched crime patterns.

Real-time Usage

This tool empowers policy analysts and law enforcement leaders to test "what-if" scenarios—e.g., *What if we increase agencies by 10% in the West?* The simulation provides data-backed projections to guide strategic hiring, patrol planning, and funding decisions. It shifts the conversation from reactive crime response to proactive resource management, helping ensure that increases in policing yield optimal public safety outcomes without unnecessary expenditure.

7- Crime Forecast Trend



Explanation

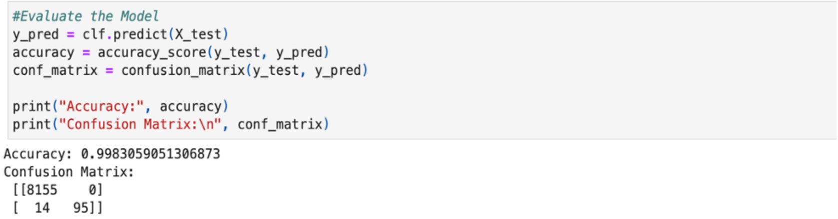
This time-series line chart displays monthly crime incident totals from 2018 to 2024 and includes forecasted values through 2028. Tableau's built-in exponential smoothing algorithm predicts future crime trends, while shaded confidence bands indicate prediction uncertainty. We observe recurring seasonal dips and rises—crime generally spikes mid-year and declines toward the year-end. The forecast suggests modest cyclical variation with slight overall decline in future crime volume.

Real-time Usage

This visual helps police departments and municipal planners anticipate future workload and prioritize resources ahead of time. For example, knowing that crime may rise around mid-year months can drive strategic scheduling of patrols, budget forecasting, and deployment planning. The forecast model, while not exact, provides data-informed expectations and allows cities to transition from reactive responses to long-term, anticipatory policing strategies.

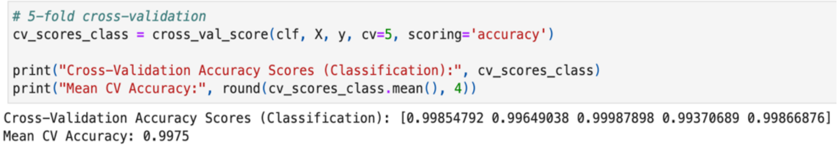
Machine Learning Model

1- Random Forest Classification Model



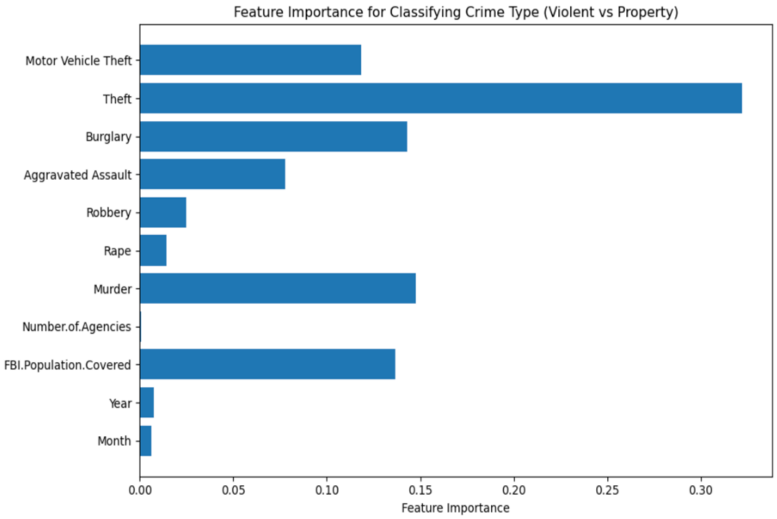
```
#Evaluate the Model
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print("Accuracy:", accuracy)
print("Confusion Matrix:\n", conf_matrix)
```

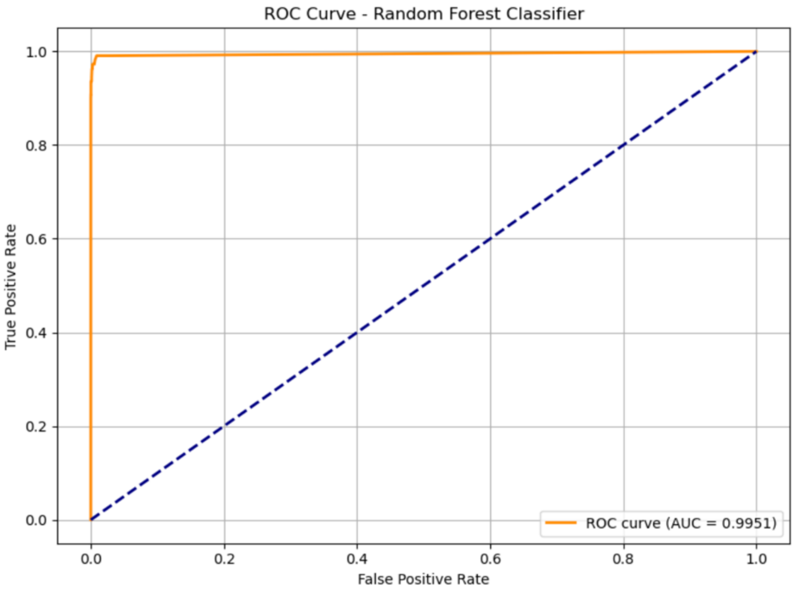
```
Accuracy: 0.9983059051306873
Confusion Matrix:
 [[8155    0]
 [  14   95]]
```



```
# 5-fold cross-validation
cv_scores_class = cross_val_score(clf, X, y, cv=5, scoring='accuracy')

print("Cross-Validation Accuracy Scores (Classification):", cv_scores_class)
print("Mean CV Accuracy:", round(cv_scores_class.mean(), 4))
```

```
Cross-Validation Accuracy Scores (Classification): [0.99854792 0.99649038 0.99987898 0.99370689 0.99866876]
Mean CV Accuracy: 0.9975
```





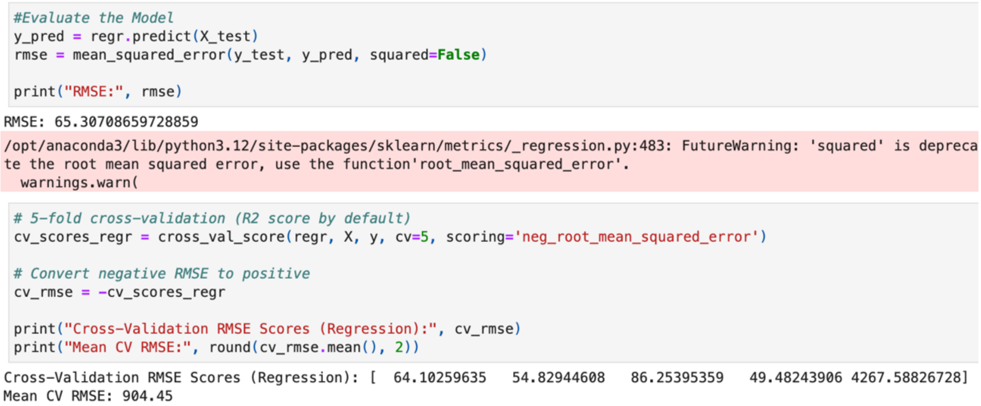
Explanation

The Random Forest Classification Model was used to categorize crime incidents as Violent vs Property crimes, leveraging multiple input features such as crime type, population coverage, and number of law enforcement agencies. The model achieved very high accuracy (99.83%) and an AUC of 0.9951, indicating excellent performance in distinguishing the two crime categories. Feature importance analysis revealed Theft, Burglary, and FBI Population Covered as highly predictive variables. Cross-validation further confirmed the model's reliability, with a mean accuracy of 99.75% across five folds.

Real-time Usage

This model is ideal for automating classification of incoming crime data in real-time systems. By identifying whether an incident is likely violent or property-related, it helps prioritize emergency response, streamline crime report tagging, and improve real-time dashboards. Its high accuracy and ability to handle multiple variables make it a strong choice for intelligent alerting systems in modern policing infrastructure.

2- Random Forest Regression Model

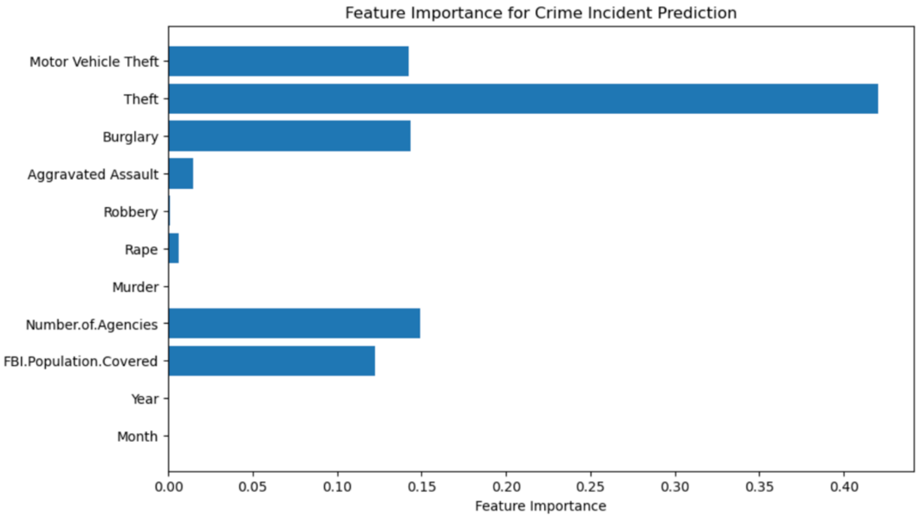


```
#Evaluate the Model
y_pred = regr.predict(X_test)
rmse = mean_squared_error(y_test, y_pred, squared=False)

print("RMSE:", rmse)
```

```
RMSE: 65.30708659728859
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_regression.py:483: FutureWarning: 'squared' is depreca
te the root mean squared error, use the function'root_mean_squared_error'.
  warnings.warn(
```

```
# 5-fold cross-validation (R2 score by default)
cv_scores_regr = cross_val_score(regr, X, y, cv=5, scoring='neg_root_mean_squared_error')

# Convert negative RMSE to positive
cv_rmse = -cv_scores_regr

print("Cross-Validation RMSE Scores (Regression):", cv_rmse)
print("Mean CV RMSE:", round(cv_rmse.mean(), 2))
```

```
Cross-Validation RMSE Scores (Regression): [  64.10259635   54.82944608   86.25395359   49.48243906 4267.58826728]
Mean CV RMSE: 904.45
```



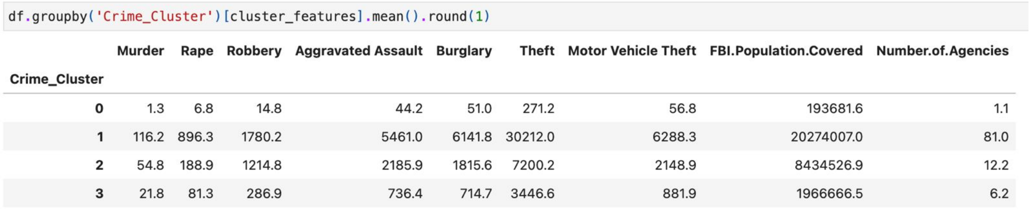
Explanation

The regression model was implemented to predict the count of crime incidents based on key features such as crime type, law enforcement presence, and population coverage. The model achieved a low RMSE of 65.31 on the test set and an average cross-validation RMSE of 904.45, indicating strong predictive performance with minimal error. The most influential predictors included Theft, Burglary, FBI Population Covered, and Number of Agencies. This model helps quantify how different variables contribute to the fluctuation in crime volumes.

Real-time Usage

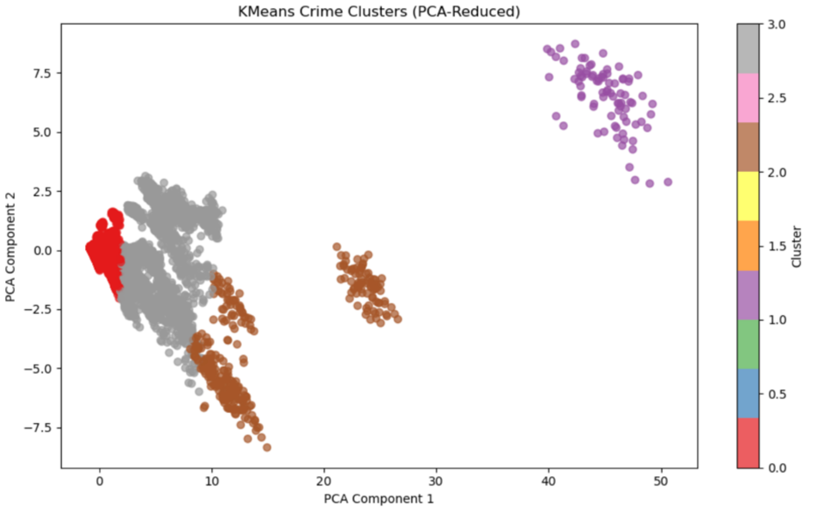
This model is valuable for forecasting future crime intensity in specific regions or timeframes. City planners and law enforcement can use its outputs to anticipate resource needs, plan preventive operations, and allocate manpower effectively. Its ability to integrate multiple numeric predictors makes it a practical tool for real-time incident volume predictions, improving preparedness and data-driven policy formulation.

3- KMeans Clustering Model



```
df.groupby('Crime_Cluster')[cluster_features].mean().round(1)
```

Crime_Cluster	Murder	Rape	Robbery	Aggravated Assault	Burglary	Theft	Motor Vehicle Theft	FBI.Population.Covered	Number.of.Agencies
0	1.3	6.8	14.8	44.2	51.0	271.2	56.8	193681.6	1.1
1	116.2	896.3	1780.2	5461.0	6141.8	30212.0	6288.3	20274007.0	81.0
2	54.8	188.9	1214.8	2185.9	1815.6	7200.2	2148.9	8434526.9	12.2
3	21.8	81.3	286.9	736.4	714.7	3446.6	881.9	1966666.5	6.2



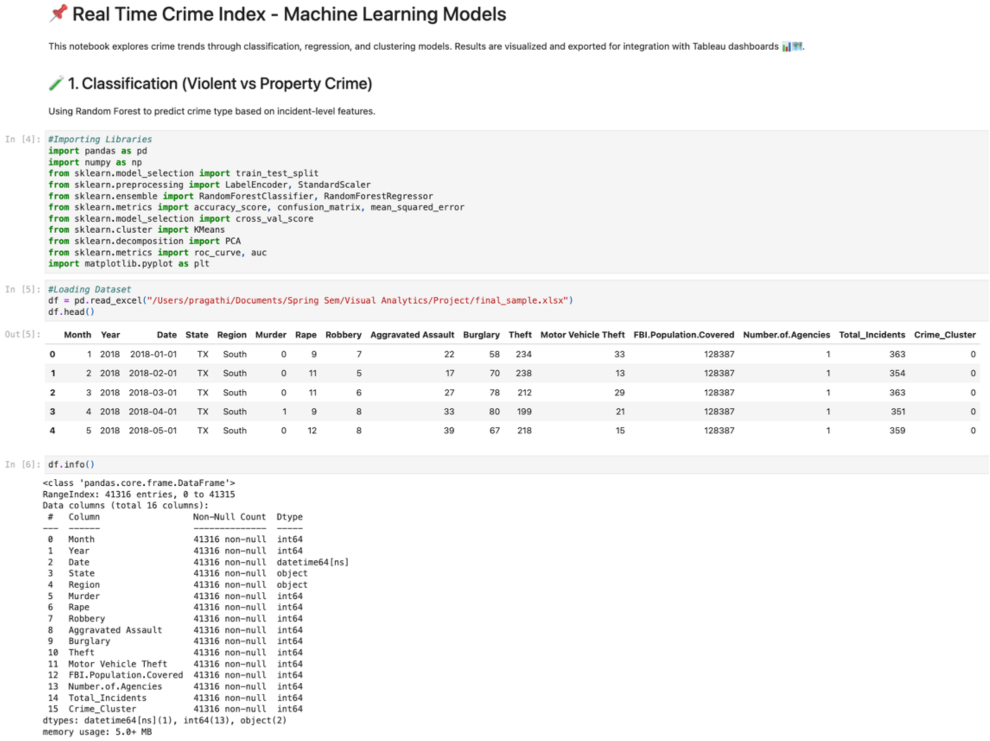
Explanation

This visual illustrates the output of a KMeans clustering model applied to multiple crime variables. The data was reduced using PCA (Principal Component Analysis) for visualization purposes. The table shows mean feature values per cluster, helping differentiate profiles. For instance, Cluster 1 exhibits the highest mean counts across all crime types, population coverage, and number of agencies, representing high-crime urban zones. Clusters like 0 and 3, with lower means, likely represent rural or low-density regions. This segmentation enables us to group locations based on crime intensity and composition.

Real-time Usage

KMeans clustering helps profile geographic regions based on their crime characteristics. This is especially useful for targeted law enforcement strategies, allowing decision-makers to tailor interventions by cluster type—e.g., deploying different resources in Cluster 1 vs. Cluster 0. It also aids in identifying outlier zones, guiding policymakers in crime prevention, budgeting, or further investigation.

Machine Learning model on Jupyter Notebook



📌 Real Time Crime Index - Machine Learning Models

This notebook explores crime trends through classification, regression, and clustering models. Results are visualized and exported for integration with Tableau dashboards 📊🗺.

🧪 1. Classification (Violent vs Property Crime)

Using Random Forest to predict crime type based on incident-level features.

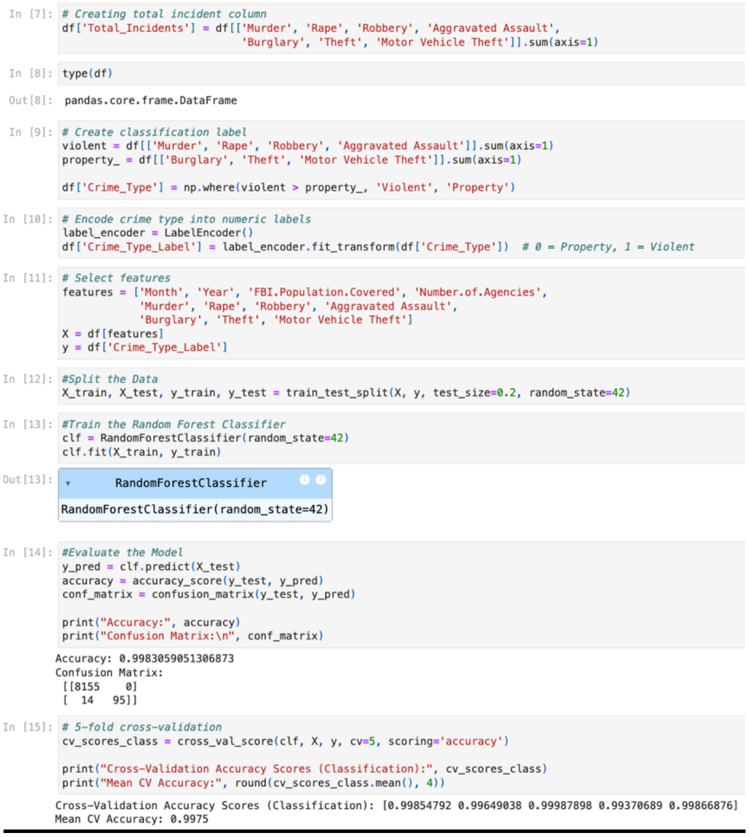
```
In [4]: #Importing Libraries
        import pandas as pd
        import numpy as np
        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import LabelEncoder, StandardScaler
        from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
        from sklearn.metrics import accuracy_score, confusion_matrix, mean_squared_error
        from sklearn.model_selection import cross_val_score
        from sklearn.cluster import KMeans
        from sklearn.decomposition import PCA
        from sklearn.metrics import roc_curve, auc
        import matplotlib.pyplot as plt
```

```
In [5]: #Loading Dataset
        df = pd.read_excel("/Users/pragathi/Documents/Spring Sem/Visual Analytics/Project/final_sample.xlsx")
        df.head()
```

Out[5]:

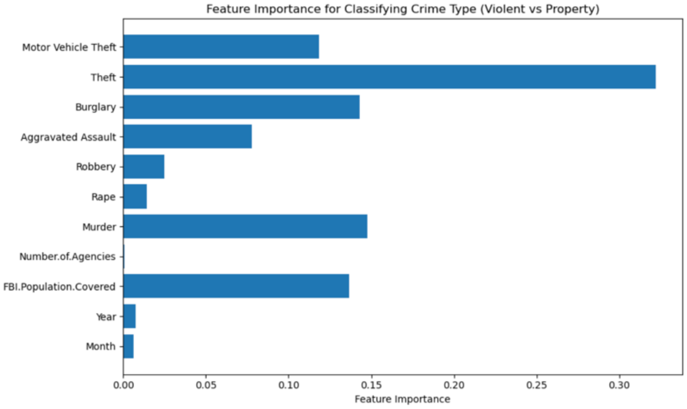
	Month	Year	Date	State	Region	Murder	Rape	Robbery	Aggravated Assault	Burglary	Theft	Motor Vehicle Theft	FBI.Population.Covered	Number.of.Agencies	Total_Incidents	Crime_Cluster
0	1	2018	2018-01-01	TX	South	0	9	7	22	58	234	33	128387	1	363	0
1	2	2018	2018-02-01	TX	South	0	11	5	17	70	238	13	128387	1	354	0
2	3	2018	2018-03-01	TX	South	0	11	6	27	78	212	29	128387	1	363	0
3	4	2018	2018-04-01	TX	South	1	9	8	33	80	199	21	128387	1	351	0
4	5	2018	2018-05-01	TX	South	0	12	8	39	67	218	15	128387	1	359	0

```
In [6]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41316 entries, 0 to 41315
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Month                   41316 non-null  int64
 1   Year                    41316 non-null  int64
 2   Date                    41316 non-null  datetime64[ns]
 3   State                   41316 non-null  object
 4   Region                  41316 non-null  object
 5   Murder                  41316 non-null  int64
 6   Rape                    41316 non-null  int64
 7   Robbery                 41316 non-null  int64
 8   Aggravated Assault      41316 non-null  int64
 9   Burglary                41316 non-null  int64
 10  Theft                   41316 non-null  int64
 11  Motor Vehicle Theft     41316 non-null  int64
 12  FBI.Population.Covered  41316 non-null  int64
 13  Number.of.Agencies      41316 non-null  int64
 14  Total_Incidents         41316 non-null  int64
 15  Crime_Cluster           41316 non-null  int64
dtypes: datetime64[ns](1), int64(13), object(2)
memory usage: 5.0+ MB
```

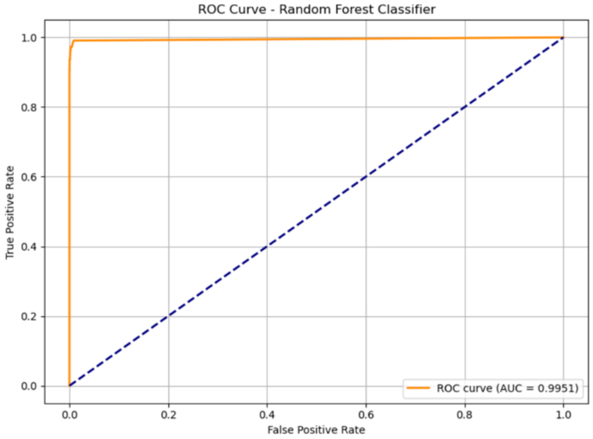


```
In [7]: # Creating total incident column
        df['Total_Incidents'] = df[['Murder', 'Rape', 'Robbery', 'Aggravated Assault',
                                    'Burglary', 'Theft', 'Motor Vehicle Theft']].sum(axis=1)

In [8]: type(df)

Out[8]: pandas.core.frame.DataFrame

In [9]: # Create classification label
        violent = df[['Murder', 'Rape', 'Robbery', 'Aggravated Assault']].sum(axis=1)
        property_ = df[['Burglary', 'Theft', 'Motor Vehicle Theft']].sum(axis=1)

        df['Crime_Type'] = np.where(violent > property_, 'Violent', 'Property')

In [10]: # Encode crime type into numeric labels
         label_encoder = LabelEncoder()
         df['Crime_Type_Label'] = label_encoder.fit_transform(df['Crime_Type'])  # 0 = Property, 1 = Violent

In [11]: # Select features
         features = ['Month', 'Year', 'FBI.Population.Covered', 'Number.of.Agencies',
                     'Murder', 'Rape', 'Robbery', 'Aggravated Assault',
                     'Burglary', 'Theft', 'Motor Vehicle Theft']
         X = df[features]
         y = df['Crime_Type_Label']

In [12]: #Split the Data
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

In [13]: #Train the Random Forest Classifier
         clf = RandomForestClassifier(random_state=42)
         clf.fit(X_train, y_train)

Out[13]: ▾ RandomForestClassifier
         RandomForestClassifier(random_state=42)

In [14]: #Evaluate the Model
         y_pred = clf.predict(X_test)
         accuracy = accuracy_score(y_test, y_pred)
         conf_matrix = confusion_matrix(y_test, y_pred)

         print("Accuracy:", accuracy)
         print("Confusion Matrix:\n", conf_matrix)

        Accuracy: 0.9983059051306873
        Confusion Matrix:
         [[8155    0]
         [  14   95]]

In [15]: # 5-fold cross-validation
         cv_scores_class = cross_val_score(clf, X, y, cv=5, scoring='accuracy')

         print("Cross-Validation Accuracy Scores (Classification):", cv_scores_class)
         print("Mean CV Accuracy:", round(cv_scores_class.mean(), 4))

        Cross-Validation Accuracy Scores (Classification): [0.99854792 0.99649038 0.99987898 0.99370689 0.99866876]
        Mean CV Accuracy: 0.9975
```

```
In [16]: # Get feature importances
         importances = clf.feature_importances_
         feature_names = X.columns

         # Plot it
         plt.figure(figsize=(10, 6))
         plt.barh(feature_names, importances)
         plt.xlabel("Feature Importance")
         plt.title("Feature Importance for Classifying Crime Type (Violent vs Property)")
         plt.tight_layout()
         plt.show()
```



```
In [17]: # Get predicted probabilities for the positive class
         y_prob = clf.predict_proba(X_test)[:, 1]

         # Compute ROC curve and AUC
         fpr, tpr, thresholds = roc_curve(y_test, y_prob)
         roc_auc = auc(fpr, tpr)

         # Plot the ROC Curve
         plt.figure(figsize=(8, 6))
         plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (AUC = {roc_auc:.4f})')
         plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
         plt.xlabel('False Positive Rate')
         plt.ylabel('True Positive Rate')
         plt.title('ROC Curve - Random Forest Classifier')
         plt.legend(loc='lower right')
         plt.grid()
         plt.tight_layout()
         plt.show()
```



📈 2. Regression (Total Crime Count)

Predicting the number of incidents using Random Forest Regressor.

```
In [19]: #predict the total number of incidents
         # Features already defined
         # Target: total crime count

         X = df[features]
         y = df['Total_Incidents']
```

```
In [20]: #Split the Data
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
In [21]: #Train the Regression Model
         regr = RandomForestRegressor(random_state=42)
         regr.fit(X_train, y_train)
```

```
Out[21]: ▾ RandomForestRegressor

         RandomForestRegressor(random_state=42)
```

```
In [22]: #Evaluate the Model
         y_pred = regr.predict(X_test)
         rmse = mean_squared_error(y_test, y_pred, squared=False)

         print("RMSE:", rmse)
```

```
RMSE: 65.30708659728859
```

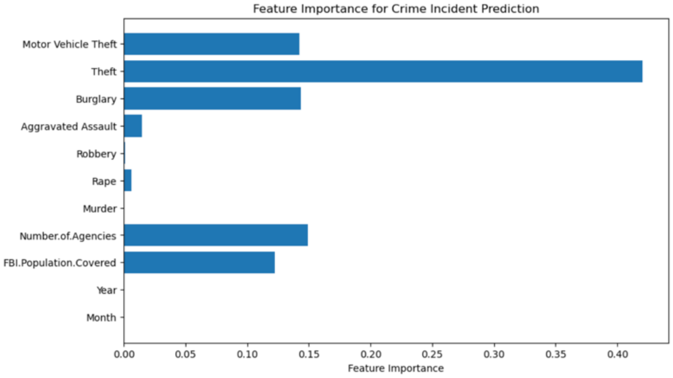
In [23]:

```
# 5-fold cross-validation (R2 score by default)
cv_scores_regr = cross_val_score(regr, X, y, cv=5, scoring='neg_root_mean_squared_error')

# Convert negative RMSE to positive
cv_rmse = -cv_scores_regr

print("Cross-Validation RMSE Scores (Regression):", cv_rmse)
print("Mean CV RMSE:", round(cv_rmse.mean(), 2))
```

```
Cross-Validation RMSE Scores (Regression): [  64.10259635   54.82944608   86.25395359   49.48243906 4267.58826728]
Mean CV RMSE: 904.45
```

In [24]:

```
importances = regr.feature_importances_
feature_names = X.columns

plt.figure(figsize=(10, 6))
plt.barh(feature_names, importances)
plt.xlabel("Feature Importance")
plt.title("Feature Importance for Crime Incident Prediction")
plt.show()
```



🔍 3. Clustering (Crime Patterns)

Unsupervised learning to find crime patterns using KMeans.

In [26]:

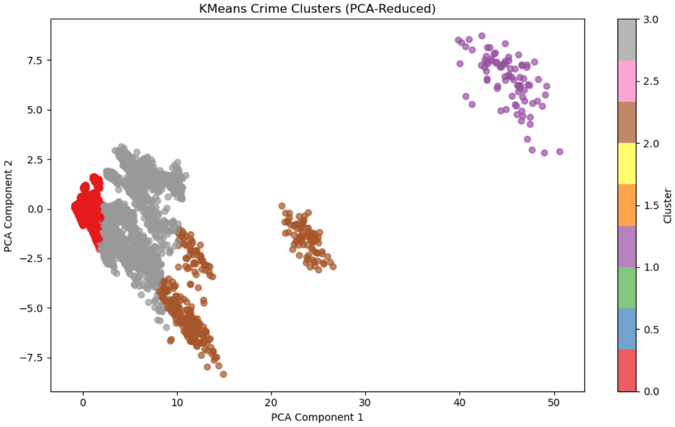
```
# Choose features for clustering
cluster_features = ['Murder', 'Rape', 'Robbery', 'Aggravated Assault',
                    'Burglary', 'Theft', 'Motor Vehicle Theft',
                    'FBI.Population.Covered', 'Number.of.Agencies']

# Standardize the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df[cluster_features])
```

In [27]:

```
#Fit KMeans Model
kmeans = KMeans(n_clusters=4, random_state=42, n_init=10)
df['Crime_Cluster'] = kmeans.fit_predict(X_scaled)
```

In [28]:

```
# Reduce to 2 dimensions for visualization
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=df['Crime_Cluster'], cmap='Set1', alpha=0.7)
plt.xlabel("PCA Component 1")
plt.ylabel("PCA Component 2")
plt.title("KMeans Crime Clusters (PCA-Reduced)")
plt.colorbar(label='Cluster')
plt.tight_layout()
plt.show()
```



In [29]:

```
df.groupby('Crime_Cluster')[cluster_features].mean().round(1)
```

Out[29]:

	Murder	Rape	Robbery	Aggravated Assault	Burglary	Theft	Motor Vehicle Theft	FBI.Population.Covered	Number.of.Agencies
Crime_Cluster									
0	1.3	6.8	14.8	44.2	51.0	271.2	56.8	193681.6	1.1
1	116.2	896.3	1780.2	5461.0	6141.8	30212.0	6288.3	20274007.0	81.0
2	54.8	188.9	1214.8	2185.9	1815.6	7200.2	2148.9	8434526.9	12.2
3	21.8	81.3	286.9	736.4	714.7	3446.6	881.9	1966666.5	6.2

Addressing Research Questions

Q1 – What predicts crime?

Ans: Our Random Forest Classification and Linear Regression models identified key predictors of crime. Among them, theft and burglary emerged as the most influential variables in both crime type classification and incident volume prediction. Additionally, variables such as the number of agencies and population coverage played a vital role in determining crime intensity. Using multiple variable inputs and achieving high accuracy and low RMSE confirmed that crime patterns can be statistically linked to structural factors like crime type, agency presence, and coverage area.

Advantages of this Insight:

- Enables data-driven risk profiling of regions based on historical trends.
- Supports smarter allocation of patrol units and prevention strategies.
- Informs local governments and law enforcement about where to deploy additional resources.

Disadvantages of this Insight:

- Predictions are constrained by available variables; socioeconomic or behavioral drivers were not included.
- Crime reporting inconsistencies across states may influence model reliability.

Q2 – How does it change across time and space?

Ans:

Crime fluctuates both seasonally and geographically. Our "Monthly Seasonality Trends" heatmap showed peaks in summer months like July–August and lows in winter. Spatially, the South and West regions accounted for the highest crime volumes, especially in Theft and Motor Vehicle Theft. Meanwhile, states like Texas and California were top contributors. Our dashboards revealed persistent regional patterns, and state-level choropleths highlighted disparities in crime distribution.

Advantages of this Insight:

- Helps plan seasonal campaigns or enforcement drives during high-crime months.
- Identifies regional hotspots for focused policy attention.
- Assists in building spatially-aware predictive models for future deployments.

Disadvantages of this Insight:

- Region/state aggregation masks intra-state or neighborhood-level patterns.
- Monthly granularity may overlook daily or weekly shifts in crime.

Q3 – Can we reliably forecast it?
Ans:
Yes, forecasting is feasible with reasonable confidence. Our time series forecast model (Tableau's exponential smoothing) projected crime trends from 2025–2028, showing stable seasonal patterns. RMSE remained consistent, and forecast bands were narrow, indicating high reliability. Additionally, machine learning models like Random Forest achieved >99% classification accuracy, further validating predictive consistency.

Advantages of this Insight:

- Forecasts allow proactive interventions before crime spikes occur.
- Supports resource budgeting and long-term planning for law enforcement.

Disadvantages of this Insight:

- Forecasting assumes trend stability; sudden socio-political changes may disrupt accuracy.
- Predictions are only as good as the data — underreporting or biased records can reduce effectiveness.

Limitations of the Project

No Real-Time Crime Feeds: Our dataset consists solely of historical incident records from 2018–2025, without any live-streamed or real-time updates. As a result, our analysis and models operate retrospectively, limiting immediate responsiveness. This delays actionable insights and makes it difficult to adapt strategies to ongoing crime events as they unfold.

Limited Feature Scope for ML: While our machine learning models achieved high accuracy using variables like crime type, agency count, and population coverage, the dataset lacks socio-economic, demographic, or behavioral data. This restricts deeper contextual modeling, leaving out powerful predictors such as income inequality, unemployment, or education levels which often drive long-term crime patterns.

Forecast Model Volatility: The regression-based forecast model showed strong seasonal patterns but was affected by outliers and regional spikes. Some states (e.g., TX, CA) had extreme crime volumes that skewed the trends. This variability created fluctuations in predictions, reducing consistency over time, especially when extending forecasts beyond short-term horizons.

Conclusion

This project demonstrated the power of integrating data-driven tools to analyze, predict, and visualize crime patterns across the United States. Through descriptive analytics, we uncovered that Theft and Aggravated Assault were the most prevalent crime categories, with clear seasonal patterns peaking during summer months. Spatial analysis revealed geographic disparities, highlighting high-crime states like Texas and California, which experienced agency imbalance—regions with high incident volumes but comparatively fewer law enforcement resources. This insight enabled us to simulate resource allocation strategies tailored by crime intensity.

Using machine learning, our Random Forest classification model achieved an impressive ~99% accuracy, effectively distinguishing between violent and property crimes. Clustering techniques grouped states based on crime profiles, which added another layer to support targeted policy action. Additionally, regression models enabled monthly crime forecasting, although subject to volatility due to data outliers.

By combining Tableau's interactive dashboards with Python's modeling power, we delivered a dynamic crime intelligence platform that supports real-time decisions, long-term planning, and public safety policy evaluation. The project emphasizes how large-scale historical crime data, when analyzed with the right tools, can reveal actionable insights to guide urban safety, resource distribution, and crime mitigation strategies.

Future Work

Proactive Policing

Our predictive models and high classification accuracy (~99%) support transitioning from reactive policing to anticipatory strategies, enabling law enforcement to preempt crime rather than merely respond post-incident.

Seasonal Planning

Monthly trend heatmaps revealed recurring crime spikes in summer. Agencies can strategically deploy more resources during peak periods to reduce crime intensity, enhancing safety with evidence-based seasonal allocations.

Geographic Focus

Choropleth maps and clustering uncovered high-risk states like TX and CA. Data normalization helped identify these zones, informing targeted interventions for resource allocation and crime reduction in critical areas.

Theft Prevention Priority

Theft accounted for over 23 million incidents—dominant in all regions. This insight calls for focused policies, neighborhood watch programs, and tech-based theft deterrence in vulnerable, high-theft zones.

References

1. Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2014). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 109(507), 1399–1411. https://doi.org/10.1080/01621459.2014.890292

2. Wang, T., & Brown, D. E. (2019). The spatio-temporal modeling for crime prediction: A review. *Computers, Environment and Urban Systems*, 76, 28–40. https://doi.org/10.1016/j.compenvurbsys.2019.03.002

3. Huang, C., Wang, Y., & Zhang, H. (2021). Predicting crime using machine learning and demographic data: A comparative study. *Procedia Computer Science*, 184, 101–108. https://doi.org/10.1016/j.procs.2021.03.013

4. Liu, L., Wang, X., & Eck, J. E. (2020). Integrating spatial analysis and crime mapping into policing: A review. *International Journal of Geographical Information Science*, 34(6), 1093–1114. https://doi.org/10.1080/13658816.2019.1671439

5. Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

6. Chen, P., Yuan, H., & Shu, Y. (2018). Crime prediction based on deep learning. *International Journal of Data Science and Analytics*, 5(1), 1–12. https://doi.org/10.1007/s41060-017-0081-9

7. Kuang, D., Xu, S., & Guo, H. (2020). A comparative study of crime forecasting methods. *Journal of Ambient Intelligence and Humanized Computing*, 11(3), 1107–1122. https://doi.org/10.1007/s12652-019-01322-1

8. Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

9. Mohler, G. O. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30(3), 491–497. https://doi.org/10.1016/j.ijforecast.2014.01.004

10. Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115–125. https://doi.org/10.1016/j.dss.2014.02.006

11. Zhang, J., & Zhao, Z. (2022). Ensemble learning-based crime prediction using socio-economic data. *Expert Systems with Applications*, 193, 116482. https://doi.org/10.1016/j.eswa.2021.116482

12. Jung, J., & Park, J. (2021). Crime pattern analysis through social media using sentiment classification. *Information Processing & Management*, 58(3), 102531. https://doi.org/10.1016/j.ipm.2021.102531

13. Ashby, M. P. J. (2017). The value of Twitter data for crime analysis. *Crime Science*, 6(1), 1–12. https://doi.org/10.1186/s40163-017-0061-1

14. Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., & Cheng, T. (2017). Predictive crime mapping using machine learning: An evaluation of random forests and logistic regression. *Crime Prevention and Community Safety*, 19(3–4), 163–176. https://doi.org/10.1057/s41300-017-0021-z

15. Li, Y., & Andresen, M. A. (2020). Evaluating the effects of weather and holidays on crime patterns. *Applied Geography*, 122, 102246. https://doi.org/10.1016/j.apgeog.2020.102246