## 1. Project Aim

The primary aim of this project is to assess and predict factors affecting crime rates in Chicago by analyzing relationships between crime incidence and variables such as weather conditions, demographics, location, and temporal trends. The ultimate goal is to identify environmental or structural drivers of crime and generate a model capable of anticipating changes in crime volume across different community areas.

## 2. Machine Learning Methodology

The machine learning workflow incorporated the Gradient Boosting Regressor from scikit-learn. The model was trained to predict crime rates per 100,000 people using a combination of historical crime data, weather statistics (e.g., temperature, humidity), and derived temporal features (e.g., day of the week, holidays, and daylight hours). Lag features were included to capture temporal dependencies.

## 3. Model Evaluation & Results

The Gradient Boosting model achieved a moderate $R^2$ score (approx. 0.30), indicating that it captured some—but not all—of the variance in the crime rate data. The MAE (Mean Absolute Error) was used to interpret prediction accuracy in real-world terms. Although performance was acceptable, it suggests room for improvement.

Prediction Example:
- For May 12, 2025, the model predicted crime counts for various community areas. While close in some regions, discrepancies highlight potential weaknesses in overfitting, feature representation, or limited data granularity.

The moderate $R^2$ reflects the complexity of crime as a social phenomenon—likely influenced by unmeasured variables such as policing patterns, local events, or economic stressors.

## 4. Limitations and Data Issues

- Missing values were handled with default imputation strategies (fillna), which may introduce bias.
- Lag features (lag_1, lag_7) can propagate errors and assume temporal continuity, which may not always hold.
- External factors (e.g., public events, local policies) not included in the dataset likely influenced crime rates.
- Weather data was averaged daily, potentially smoothing out critical intra-day variations.

## 5. Critical Review Based on Hypotheses

**H1 (Weather Impact):** Supported. Correlation and regression models found a significant link between higher temperatures and increased crime. Rainfall appears to reduce total crime but changes the distribution.

**H2 (Demographics & Location):** Partially supported. Location (longitude) showed moderate spatial correlation, but demographic variables like income and race lacked strong predictive power in this ML model, despite spatial clustering observed visually.

**H3 (Crime Type Sensitivity):** Supported in earlier statistical tests, though ML predictions were made on aggregate crime rate, not broken down by type—this could be improved in future iterations.

## 6. Recommendations

- Incorporate more detailed, sub-daily weather metrics.
- Explore classification models by crime type to reflect hypothesis 3 more directly.
- Use cross-validation and hyperparameter optimization (e.g., GridSearchCV).
- Try interpretable models (e.g., SHAP with XGBoost) to quantify feature importance.
- Evaluate time-series-specific models like SARIMA or Prophet.

## 7. Conclusion

The current ML model demonstrates foundational predictive capacity for understanding how environmental and contextual variables affect crime rates. However, model performance and generalizability can be significantly improved by enhancing feature richness, capturing real-time dynamics, and incorporating socioeconomic complexity.