Author: Sheikh Khairul Momin Mohammad Tahmid
Date: 23 November 2025

Montgomery County Crime Analytics Project

Data preparation

The work starts from the raw Montgomery County incident file, which is cleaned and standardised into a single table called df_clean. The cleaning step harmonises column names, fixes date fields, constructs duration and reporting delay, and validates latitude and longitude against county bounds. After removing duplicates and unusable rows, the final dataset contains 280,708 incidents with unique incident identifiers and consistent geography and time information.

Idea 1: Crime hotspots and simple risk scores

Using this cleaned table, incidents are grouped by police district (or by city or place when district is missing). For each area the analysis counts total incidents and those that match a simple “violent” keyword rule. In the current run, the rule does not flag any incidents as violent, so risk scores are driven by volume alone. Silver Spring, Wheaton and Montgomery Village appear at the top of the ranking, while places like Gaithersburg and the residual “Other” category sit at the bottom. A scatter plot of more than 200,000 points with valid coordinates then shows how these incidents are spread across the county, highlighting dense corridors of activity.

Idea 2: Temporal patterns of crime

For time series analysis, the code builds a calendar view with year, month, weekday and hour for every incident that has a valid start time. Aggregation by month produces a 74-month series from July 2016 to August 2022. Most months lie between about 3,000 and 4,500 incidents, with a noticeable dip around early 2020 and a very low final point that reflects a partially observed month. Crime Against Property dominates the monthly breakdown, while Crime Against Person, Crime Against Society, Other and Not a Crime form smaller but stable components. Weekday plots show that Friday is the busiest day and Sunday the quietest, and the hour-of-day plots reveal a sharp peak around midnight, a quiet early morning, and a broad rise through the afternoon and evening.

Idea 3: Text based crime type classification

To model the high-level crime category, the project combines several text fields (crime_name1, crime_name2, crime_name3 and place, plus address fields where available) into a single description. After filtering, 280,708 incidents remain, and an 80/20 split produces 224,566 training samples and 56,142 test samples, each represented by 2,301 TF-IDF features. A multinomial Logistic Regression model with balanced class weights achieves a test accuracy of 1.0000. The classification report shows precision, recall and F1 of 1.00 for all five classes, and the confusion matrix has only diagonal entries, for example 26,111 Crime Against Property and 11,588 Crime Against Society incidents correctly predicted with no cross-class errors. Inspection of the learned coefficients confirms that the model is picking up highly specific terms like “larceny”, “narcotic” or “runaway juvenile”, which directly signal the target label.

Idea 4: Predicting crime severity from tabular features

The next stage defines a binary severity label: incidents in the “Crime Against Person” category are coded as severe (1), and all others as non-severe (0). The same 280,708 incidents are used, with a severe rate of about 9.5 percent. Features include numeric fields such as number of victims, offence code, latitude, longitude, duration, report delay and time of day, along with categorical variables for district, agency, place, city and NIBRS code. A stratified split again yields 224,566 training and 56,142 test rows.

Two models are trained within a shared preprocessing pipeline. Logistic Regression reaches an accuracy of about 0.998 and a ROC AUC just under 1.0. On the test set it correctly identifies 50,754 of 50,828 non-severe incidents and 5,295 of 5,314 severe incidents, missing only 19 severe cases and mislabelling 74 non-severe cases as severe. A Random Forest with 200 trees performs even better, correctly classifying all 50,828 non-severe cases and 5,291 severe cases, with only 23 severe incidents missed, and achieves an almost perfect ROC AUC. Feature importance plots show that certain NIBRS codes and the numeric offence code are the strongest predictors, with reporting delay, duration, victim count and some place indicators providing additional signal.

Idea 5: Clustering and unsupervised pattern discovery

For unsupervised analysis, the project selects seven numeric features: latitude, longitude, victims, duration, report delay, hour and day of week. After removing rows without coordinates and filling remaining gaps with medians, 274,089 incidents enter the clustering stage. A standardised KMeans model with six clusters is fitted. The resulting clusters vary widely in size, from about 5,457 incidents in the smallest group to 109,340 in the largest, with intermediate clusters of roughly 87,917, 55,312, 6,005 and 10,058 incidents.

Cluster profiles show distinct patterns. One cluster has slightly more than two victims on average, while others are close to one. Some clusters have short average reporting delays and durations; others exhibit long delays and extended incidents. The map of cluster labels across latitude and longitude reveals that each cluster occupies different parts of the county, often following transport corridors or dense urban pockets. Temporal profiles by hour of day show that one cluster peaks around midnight, while others are more active in afternoon and early evening periods, indicating different underlying crime settings.

Idea 6: Crime counts forecasting

To explore forecasting, the code aggregates incidents per month for the whole county, again producing 74 monthly counts. The first 62 months, from July 2016 to August 2021, form the training period, and the last 12 months form the test period. Two simple baselines are evaluated. The naive forecast, which repeats the last training value, yields a test RMSE of about 870.8 incidents. The seasonal naive forecast, which repeats the value from the same month one year earlier, achieves an RMSE of about 886.1.

A seasonal ARIMA model with yearly seasonality is then fitted. On the same test period it obtains an RMSE of about 906.7, slightly worse than the naive benchmarks. When refitted on the full series and used to predict the next 12 months, the SARIMA model produces wide confidence intervals and some negative forecasts, reflecting the influence of the unusually low final month in the historical data. The same workflow is repeated for the three highest volume districts, Silver Spring, Wheaton and Montgomery Village. For these districts, naive baselines again match or outperform SARIMA, with RMSE values around 150 to 200 incidents per month, which underlines the importance of comparing complex models to simple, transparent alternatives.

Idea 7: Interactive dashboard

Finally, the project bundles several outputs into a Plotly Dash dashboard. The app uses df_clean as its backbone, creates helper time columns, and exposes filters for date range, crime type and district. Three headline indicators are displayed for the filtered data: total incidents, share of severe incidents and average incidents per month. A line chart presents the monthly trend over the selected period, a bar chart ranks districts by incident count, and a Mapbox scatter plot shows up to 10,000 sampled incidents coloured by crime type with rich hover tooltips for district, place, start time, victim count and severity. The dashboard therefore provides an end-to-end view of crime volume, spatial distribution and severity in an interactive form that mirrors tools used in real analytical teams.

Overall reflection

Taken together, the notebook demonstrates a complete analytics pipeline for police incident data. It moves from careful cleaning and feature engineering, through exploratory mapping and time series visualisation, to supervised learning, unsupervised clustering, forecasting and interactive communication. The numeric results show that text-based models can reproduce existing crime type labels almost perfectly, that severity can be predicted with extremely high accuracy using structured fields, and that naive forecasting methods remain hard to beat for counts that are affected by structural breaks and incomplete months. The dashboard ties these strands together and offers a practical way for non-technical users to explore crime patterns across Montgomery County.