Team 4 – The Santa Clara Crime Dataset

Abstract

This paper presents an in-depth analysis of the Santa Clara Crime Dataset, which encompasses 194,865 police response occurrences in Santa Clara County, California, recorded from August 1, 2017, to December 15, 2019. Our team, comprising data science and research professionals, aims to contribute to public policy discussions by investigating whether crime rates in Santa Clara County have been increasing and to optimize the deployment of public safety resources. The dataset includes several types of crime incidents, and our analysis focuses on understanding temporal patterns and forecasting future occurrences. For this project, we analyzed and statistically modeled a total of sixteen time series. The statistical methods that we employed were data transformation to stationarity, Autoregressive Integrated Moving Average with Seasonality, ARMA and SARIMA, and Generalized Auto Regressive Conditional Heteroskedasticity GARCH.

Group Members

Katie Hill – Distance M.S. student, a research scientist at the Pantex plant in Amarillo, TX, is interested in data analysis and writing.
Erin Batta – In-person M.S student, currently working at the Texas A&M Public Policy Research Institute in College Station, TX, is interested in data visualization and writing.
Enrique Ortizmata – Distance M.S. student, Intelligence Operations Manager, Navy Region Northwest, is interested in data visualization and data driven decision process.
Emanuela Ene – Distance M.S. student in Statistical Data Science, a professor at Lone Star College, Houston, TX, is interested in refreshing her data analysis and multigenerational-team collaboration skills.
Case Herndon – Distance M.S. student, a research Analyst at Baylor University, is interested in data models and visualization.

Project Motivation

There has been a call to action to reform the public policies in the Bay Area, particularly in criminal justice. The crime rate has been in the national and local news with a bias favoring a "tough on crime" approach. We are personally motivated because one team member is an expert in public policies, and the family of another member is working in high-tech and living in Santa Clara County. Additionally, Seattle and other high-growth urban areas are experiencing a perception of higher-than-normal crime rates. That may be fueled by news correspondents.

Project Goal

Our team will attempt to make a positive contribution to the public discussion by deploying a responsible analysis of the data to answer the question: "Is crime really on the rise?" We will analyze the trends and areas of the most frequent types of crimes recorded in order to help optimize the use of public safety resources. We hope to build a model that will forecast future crime statistics on incident type.

Santa Clara County Background

Amid struggles with scandal and overwhelming caseload, Santa Clara County has recently been seeking to adopt new practices to improve their response to crime. Measures such as adding new devices in addition to ankle monitors[1], as well as installing more speeding cameras throughout San Jose[2], are in progress. News agencies report that there is still a high number of cases that have not yet been disposed since COVID-19.[3]

---

[1] Santa Clara County approves 'less invasive' house arrest monitors - San José Spotlight (sanjosespotlight.com)
[2] San Jose to Install First Speeding Cameras in Late 2025 | San Jose Inside
[3] Enforcement Activity 2019.pdf (sccgov.org)

While our dataset covers a slightly older span of dates (2017-2019), it can help shed light on this issue throughout time and reveal what was occurring regarding crime immediately prior to COVID-19. This provides motivation for our first research question, which was whether crime is increasing over the period we have data for.

The Santa Clara County Crime Dataset Description

This dataset [4] contains 194,865 observations of unique police response occurrences in Santa Clara County, California. It has 20 descriptor fields, including date, time of the incident, type, and location of incident. There are 27 types of primary crime incidents recorded from August 1st, 2017, to December 15th, 2019, for 867 calendar days. There are five one-day gaps and one five-day gap, between Jan.30 and Feb.5, 2018, in these recordings.

Figure 1 shows the prevalence of each primary category of crime incidents recorded in this dataset and how the total number of crimes fluctuates over the time of each day. Figure 2 shows the time evolution in monthly, respectively weekly counts.

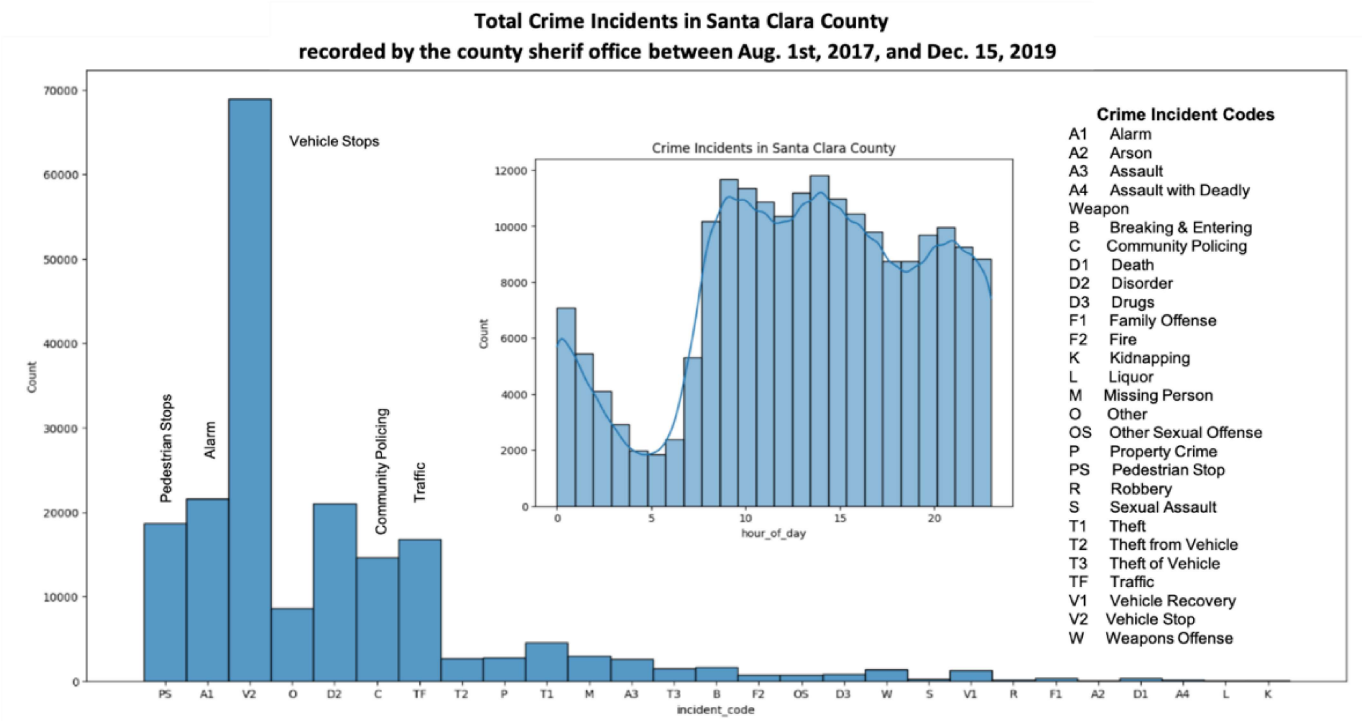

Figure 1 - Characterization of the Santa Clara Crime Dataset– Histogram of the total counts of the 27 primary crime incidents and their codes. In the inset, the hourly variation of the crime incident count in a typical day at the sheriff's office in Santa Clara County.

Exploratory Data Analysis

This dataset is a record of police responses over time, including mostly common routine checkouts but also serious crimes followed by court filing. We analyzed the occurrences of each of the 27 parent incident types. For each type, we aggregated the counts monthly, weekly, daily, and hourly. Through count aggregation, we obtained 28 time series of monthly counts, 28 time series of weekly counts, 28 time series of daily counts, and 28 time series of hourly counts, a total of 112 time series. Our analyses focused on the overall crime incident counts and for six most frequent incident types, which together accounted for 83% of the entries in the sheriff's recordings. For this project, we analyzed a total

[4] https://www.kaggle.com/datasets/vaghefi/santa-clara-country-crime

of 16 time series. Figure 2 shows two of the time series that we analyzed: the overall crime incidents summed monthly, respectively weekly.

The incident recording starts at 1 AM on Tuesday, Aug. 1st and stops at 14 PM on a Sunday, Dec. 15th. The last day, the last week, and the last months recorded in the dataset were shorter than the rest. This explains why the plots in ***Figure 2*** fall sharply for the last month and week. On the plot at the bottom, the first big dip corresponds to the 7-day period from Jan. 30, 2018, to Feb. 5, 2018. That week has 5 missing days of recordings. The final week line drop is because of the missing Sunday and Monday records. The other dips in the 7-day plot are related to the other 5 days of missing data.

Based on the gravity of the daily crime incidents, we further grouped them into four large categories: (i) common public safety crimes and traffic related; (ii) theft, family, and property crimes; (iii) drugs, liquor, and sex crimes; (iv) serious, less common crimes. ***Figure 3*** illustrates the monthly trends for four common crime incidents in category (i) compared to the monthly trends of four serious crime incidents in category (iv). The scale difference is 20:1. It is obvious that the general trend of the data is dictated by common incident types (i), which are not serious crimes, with an average of more than 12 counts per day that accounted for 83% of the total data. Four common incidents had a seven-day cycle: V2 "Vehicle stop", C "Community Policing", TF "Traffic", and D2 "Disorder". Two common incidents had an almost flat daily count pattern: A1 "Alarm" and PS "Pedestrian Stop".

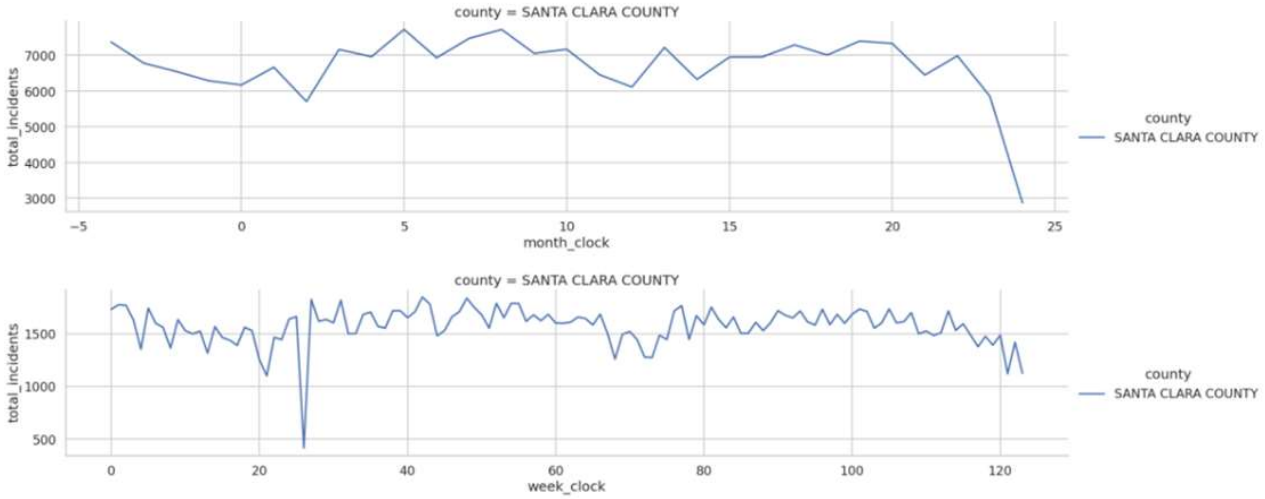

Figure 2 - Temporal evolution of the crime incidents recorded in Santa Clara County. Upper panel - Monthly counts. Lower panel - Weekly counts, from Aug. 2017 to Dec. 2019

There were 11 primary incidents with a seven-day cycle, accounting for a total out of 67% of the total data. The most frequent primary category of incidents was V1 "Vehicle stop" with 68,912 observations, accounting for 35.4% of the total data. Vehicle stops had a weekly cyclicity with a peak for midweek and, also, an hourly periodicity throughout a typical day with a dip for the early morning. During the 857 days of recording, only 14 vehicle stops were followed by filling with the court, which represents 0.02% of the total stops. It could be inferred that most vehicle stops were preventive police actions or minor incidents. Other frequent primary incidents with a weekly pattern were "Disorder" (10.8%), "Traffic" (8.6%), and "Community Policing" (7.5%). Seven cyclic incidents had less than 2 daily counts on average: "Death", "Drugs", "Other Sexual Offence", "Property Crime", "Theft from Vehicle", "Vehicle recovery", and "Weapons offence". There were 16 primary incidents with an apparently non-cyclic daily count pattern, accounting for 33% of the total data. "Alarm" plotted with 21,563 observations, accounted for 11.1% of the data, was the largest of the non-cyclic crime incidents.

Figure 4 illustrates the time evolution of the overall daily counts, the seven-day seasonal component, the overall count trend, and the random fluctuations of the daily counts. Figure 5 represents the correlograms of the full data set, and for

three relevant subsets described above. The autocorrelation function of the full set and its subsets depends only on the temporal lag and does not change in time. The dataset and its subsets are stationary.

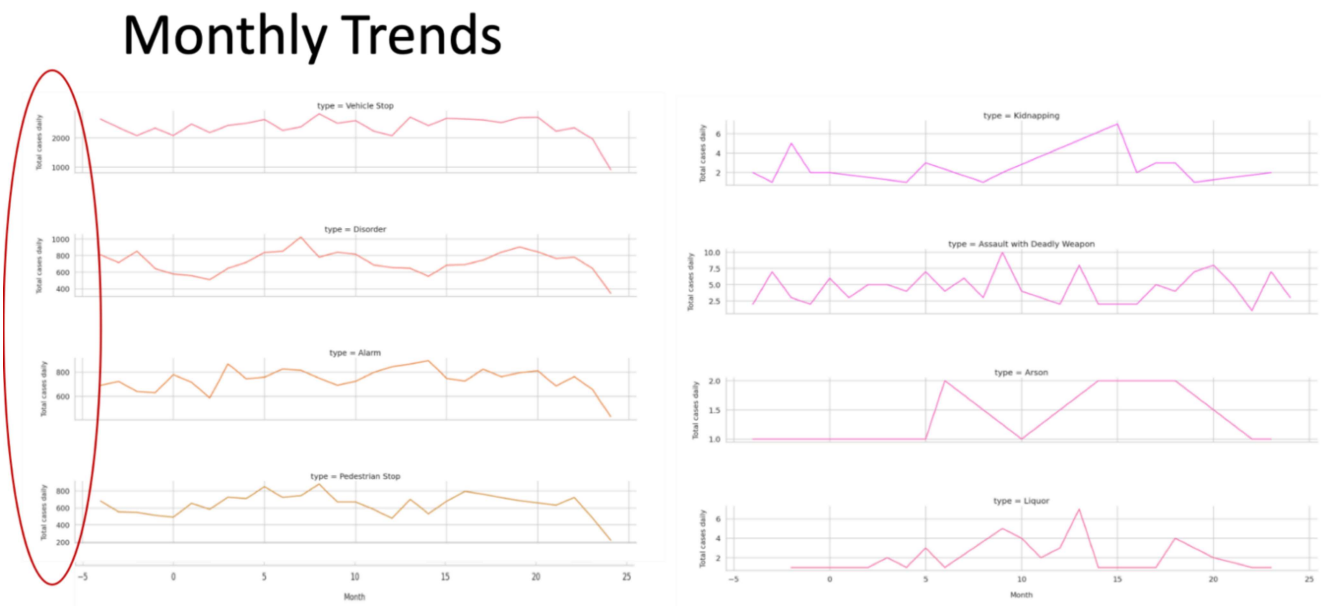

Figure 3 – Monthly evolution of crime incidents recorded in Santa Clara County. Left panel – Four common public safety crimes and traffic related types of incidents. Right panel – Four serious crime types of incidents.

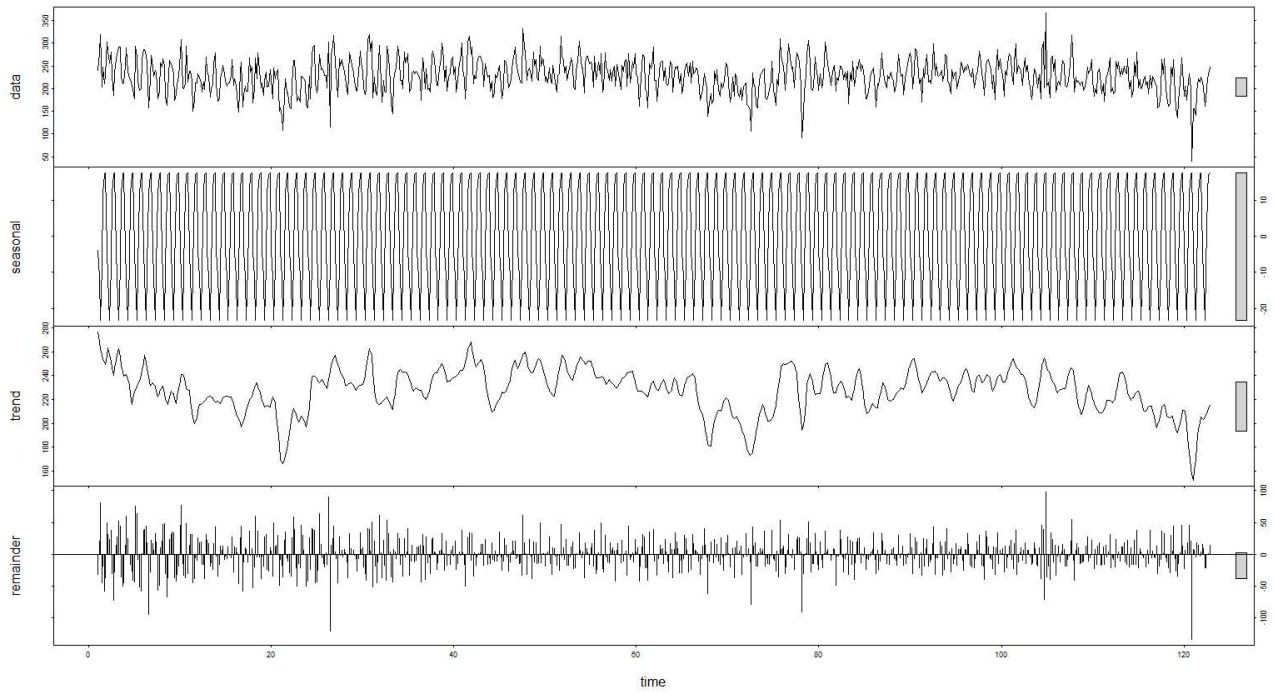

Figure 4- Daily evolution of the crime incidents recorded in Santa Clara County. Upper panel – Daily incident counts. Lower panel – data random noise after subtracting the seven-day seasonality and the overall trend illustrated in the middle panels.

It appears that crime is not increasing over time: the monthly incident sum, respectively the seven-day incident sum was constant during the 24.5 months of records. The incident counts were stationary, with 7,000 incidents/month respectively 1500 incidents/week. There were 227 daily incidents recorded in Santa Clara on average. The monthly and weekly counts were not autocorrelated, but there was a seven-day cyclicity of the daily counts.

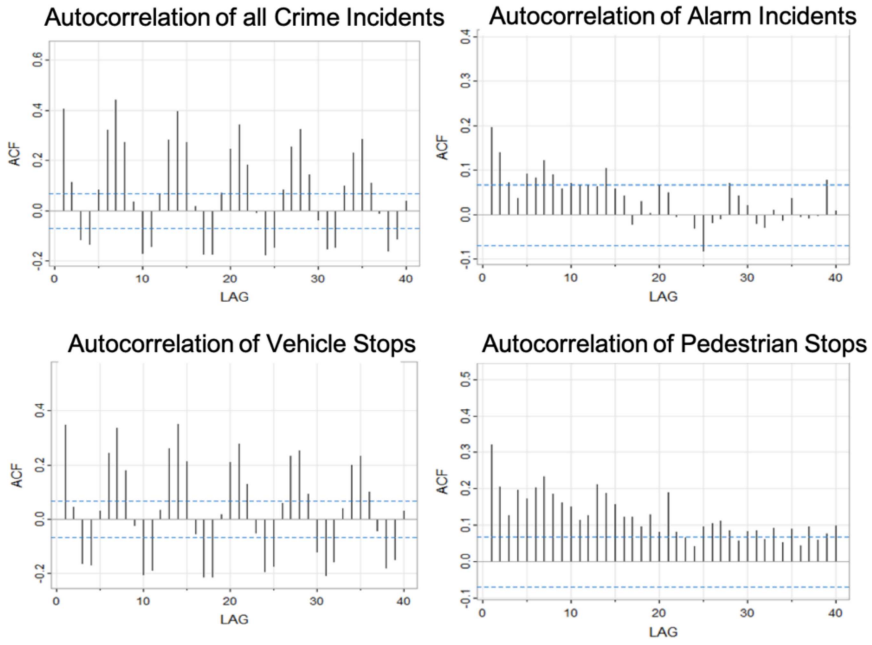

Figure 5 - Autocorrelations of the daily counts of crime incidents recorded in Santa Clara County. From upper left clockwise: the full data, the vehicle stops (35.4%), the pedestrian stops (9.6%), and the alarm incidents (11.1%).

Analysis of Weekly Counts and Forecast Plots

In Figure 6, we see periodic increases and decreases in disorder incidents over time, and we see significant autocorrelation at initial lags, suggesting that current disorder incidents are influenced by recent weeks' data. This effect diminishes over time. The PACF plot confirms significant partial correlations at the first few lags, indicating that recent weeks' incidents have a direct impact on the current week's incidents. Incidents peak in the late evening hours, particularly around 9 PM. They are also most frequent on Saturdays, with lower counts on weekdays, especially on Monday. This indicates that late evening hours, and weekend activities significantly influence disorder incidents. A model for Disorder weekly counts is ARMA (2,0,1) because the partial uncorrelation coefficients PACF cut off at lag=2week and there is no seasonality. As shown by the diagnostic plots in , the ARMA(2,0,1) model of the weekly counts yields normally distributed uncorrelated residuals, which is good, but also significant p-values for lags smaller than 10, which signals that this model is not optimal. For disorder incidents, the ARMA(2,0,1) model shows high volatility. While the standardized residuals generally behave as white noise, the ACF plot of residuals has minor lags, the Q-Q plot aligns decently well with the theoretical quantiles, and the Ljung-Box test indicates acceptable p-values, though leaving room for potential improvements, as mentioned above.

The ARMA(1,0,1) model for traffic incidents shows a stable forecast with minor fluctuations, indicating predictability in traffic-related incidents. Figure 8 reveals the diagnostic plots for traffic incidents, where the standardized residuals appear to be white noise, the ACF plot of residuals shows no significant autocorrelations, the Q-Q plot aligns closely with the theoretical quantiles, and the p-values for the Ljung-Box statistic are well above the significance level, suggesting a good model fit.

The vehicle stop incidents, also modeled with ARMA (1,0,1), exhibit more variability. The standardized residuals shown in Figure 8 do not show any significant patterns, the ACF of residuals indicates no significant lags, the Q-Q plot suggests normally distributed residuals, and the Ljung-Box statistic p-values confirm no autocorrelation in residuals.

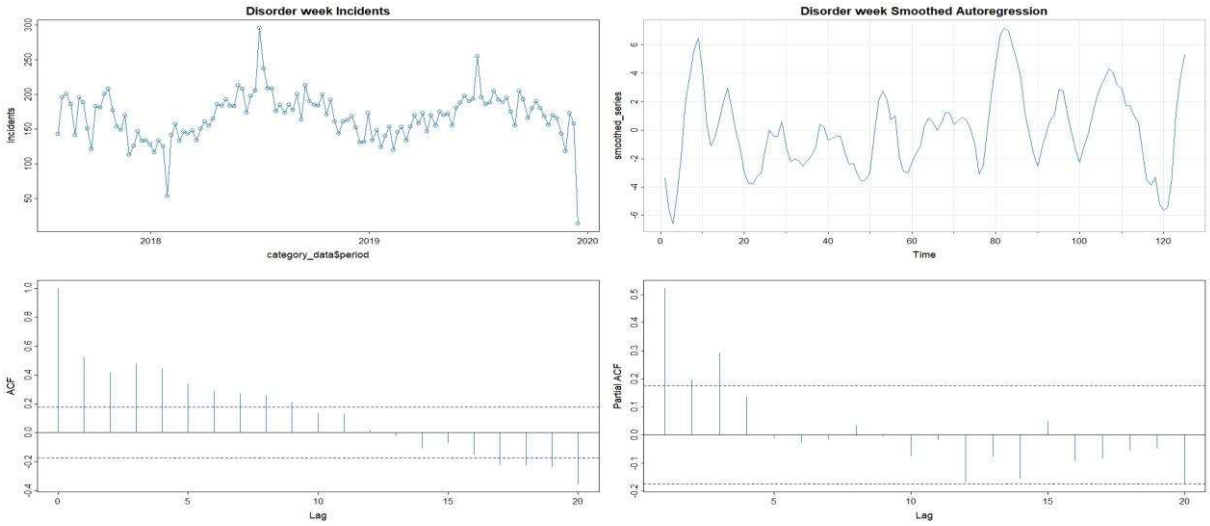

Figure 6 – *Modeling the weekly counts of the disorder incidents recorded in Santa Clara County.* *From left* to right on rows*: the full data, the smoothed weekly autoregression*, *the autocorrelation of the weekly counts, and the partial autocorrelation of the weekly counts.*

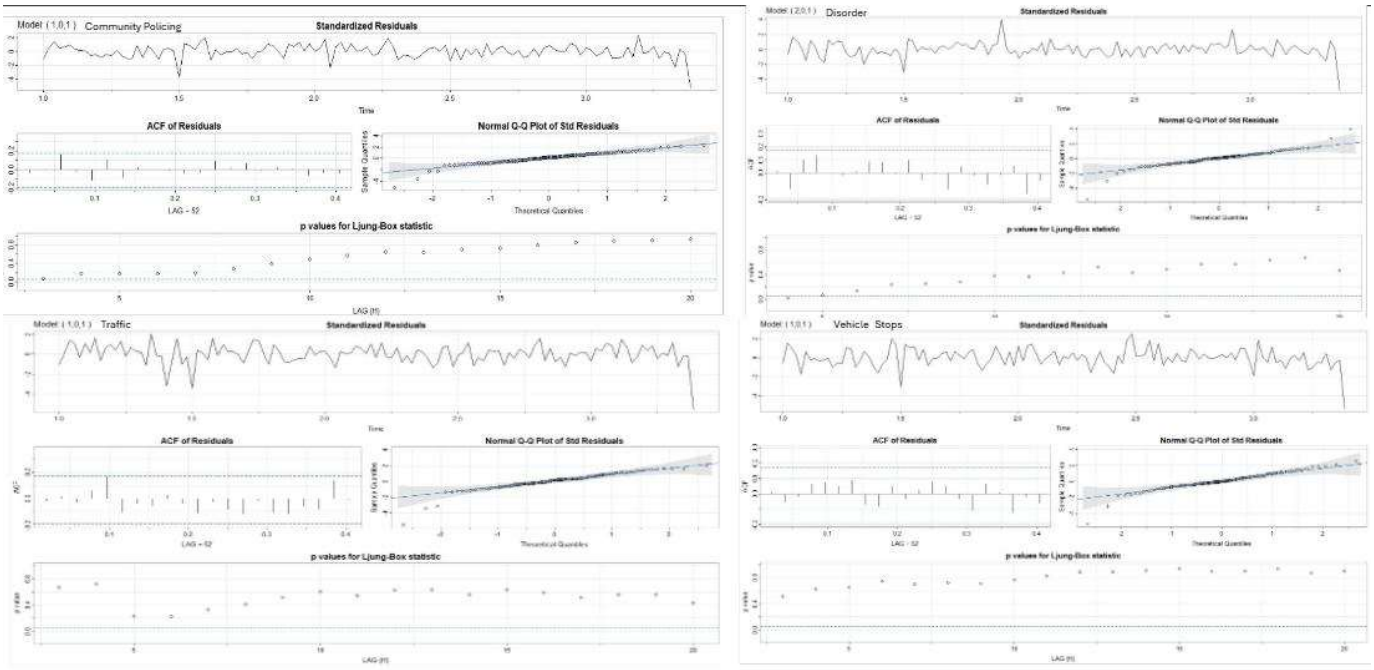

Figure 7 - Weekly Counts Model Diagnostics for Traffic, Vehicle Stop, Disorder, and Community Policing Incidents using ARMA Models.

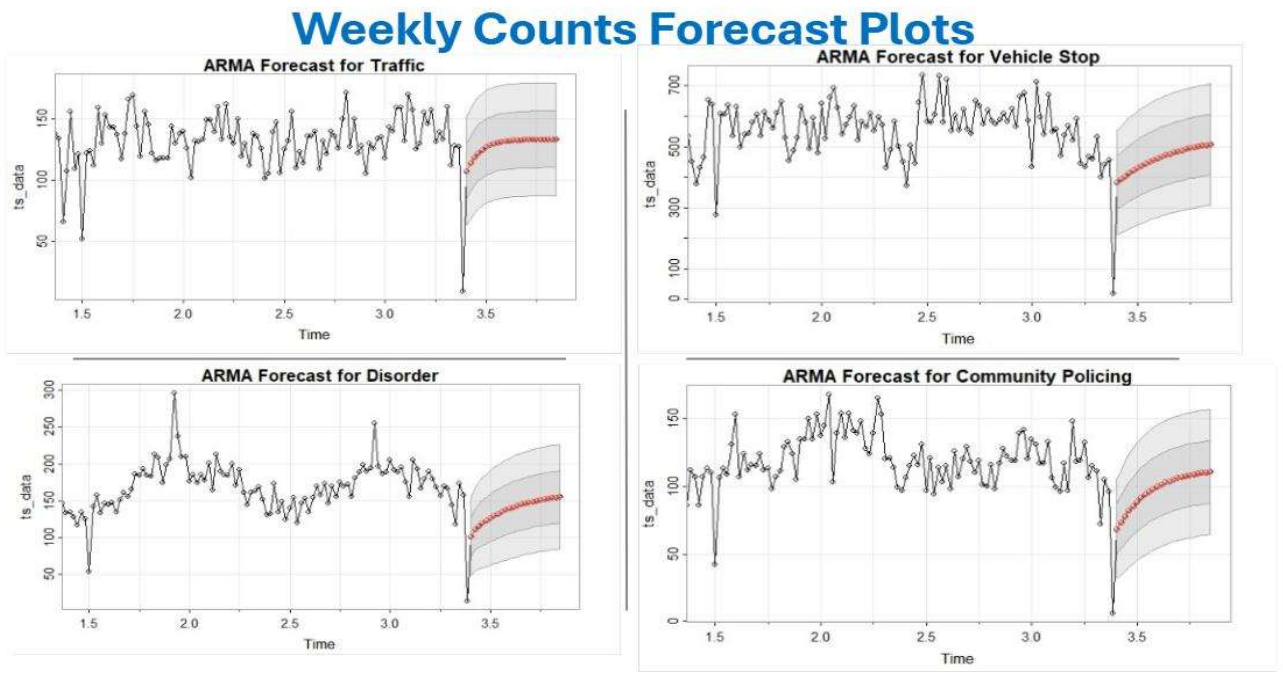

Figure 8 - Weekly Counts Forecast Plots for Traffic, Vehicle Stop, Disorder, and Community Policing Incidents using ARMA Models. The upper left panel shows the ARMA(1,1) forecast for traffic incidents, the upper right panel displays the ARMA(1,0,1) forecast for vehicle stop incidents, the lower left panel illustrates the ARMA(2,0,1) forecast for disorder incidents, and the lower right panel presents the ARMA(1,0,1) forecast for community policing incidents.

Community policing incidents, modeled with ARMA (1,0,1), display moderate fluctuations and a downward trend. In, the standardized residuals suggest white noise behavior, the ACF plot shows no significant autocorrelations, the Q-Q plot indicates normally distributed residuals, and the p-values for the Ljung-Box statistic support the absence of significant autocorrelation, indicating a good fit. Figure 8 illustrates the 26-week ahead ARMA forecast of the weekly count of four types of cyclic frequent crime discussed in this section.

Analysis of Daily Counts and Forecast Plots

For an accurate forecast of the daily counts on type of incident, we filled in the missing ten days with counts equal to the nearest similar weekday. We transformed the daily counts of the frequent crime incidents that accounted for 83% of the data for achieving data normality and time series stationarity by applying first and second order differencing, log-transforming, and nabla(log) -transforming. Supplementary to calculating the autocorrelation and partial correlation functions for each of the 27 types of parent-incidents, we ran Fast Fourier transforms of the daily counts of the six most frequent common crime incident types. The ACF, PACF, and FFT techniques revealed the seven-day seasonality but also 30-day and 365-day for community policing, vehicle stops, and disorder. We modeled the cyclic daily counts though Seasonal Autoregressive Integrated Moving Average (SARIMA) and accounted for the polychromaticity through Generalized Auto Regressive Conditional Heteroskedasticity (GARCH). The GARCH model handled the outlier and weak correlations of the residuals. The daily counts analysis was performed in R and Python.

With the noise terms conditionally modeled through GARCH, the SARIMA prediction of the daily counts of the common crime incidents during the last weeks of December 2019, were accurate. Figure 9 illustrates the forecasting for disorder, vehicle-stops, alarms, and community policing. The 95% prediction interval of two standard deviations mimics the past temporal pattern.

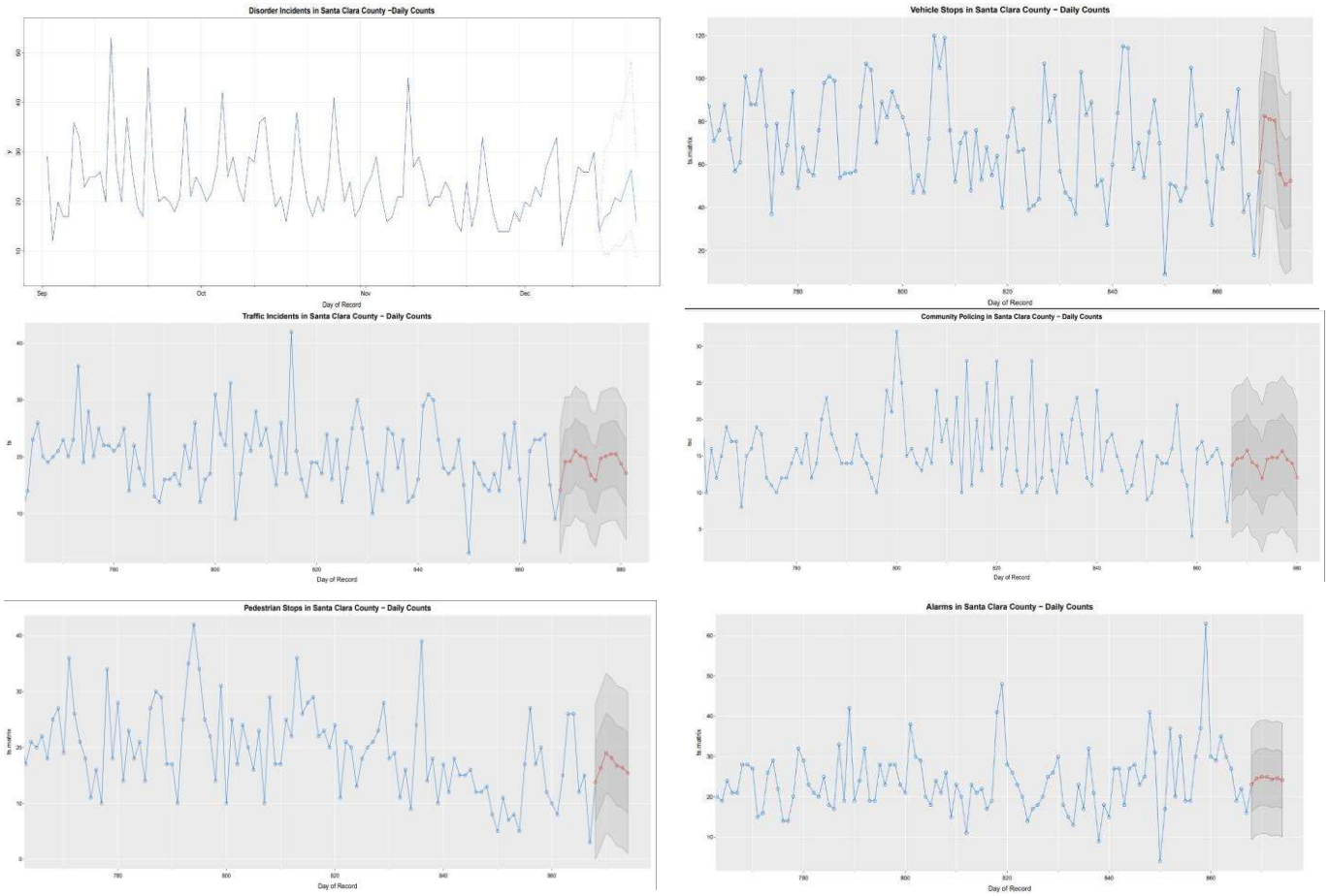

Figure 9 - ***Forecasting the daily counts after December 15, 2019.*** *From left to right on rows: (a) Disorder; (b) Vehicle Stops; (c) Traffic; (d) Community Policing; (e) Pedestrian Stops; (f) Alarms.*

We cross validated our SARIMA+GARCH estimations with the local news channels from Santa Clara County, which explained why the daily count of alarms, vehicle stops, pedestrian stops, or disorder incidents increased significantly for certain calendar days, such as the local celebrations such as 4th of July. The daily counts of disorder incidents exhibit significant variability with notable peaks and troughs on July 4th, 2018, and 2019. The diagnostic calculations reveal that the standardized residuals are relatively well-behaved, with the ACF plot showing minimal significant lags, indicating that the model is reasonably capturing the temporal dependencies. However, the Q-Q plot and the Ljung-Box statistic suggest that there may still be room for improvement in the model fit, as the residuals exhibit some degree of autocorrelation. The vehicle stops data shows more consistent fluctuations compared to disorder incidents. The diagnostics indicate that the standardized residuals follow a white noise pattern, and the ACF of residuals confirms no significant autocorrelation, which supports the model's adequacy. The Q-Q plot aligns well with the theoretical quantiles, and the Ljung-Box test results indicate a good model fit, suggesting that the model effectively captures the patterns in vehicle stop incidents. Pedestrian stop incidents display high frequency and variability. The model diagnostics show that the residuals from the ARMA model are well-distributed, with minimal autocorrelation as indicated by the ACF plot. The Q-Q plot and the Ljung-Box test confirm that the residuals are normally distributed and uncorrelated, validating the model's effectiveness. However, the significant peaks observed in the data suggest that external factors might influence these incidents, which are not fully captured by the model. The daily counts of alarms show a mix of high-frequency fluctuations and sporadic peaks. The diagnostics reveal that

while the residuals are generally well-behaved, the ACF plot indicates minor lags, and the Q-Q plot shows deviations at the tails, suggesting some non-normality. The Ljung-Box statistic further supports this observation.

Analysis of Hourly Incidents

Figure 10 and Figure 11 represent the smoothed time evolution of the hourly counts, respectively the 24-hour ahead forecast of these counts. We modeled the hourly counts on type of incident with ARMA in R and with the package PROPHET[5]. The smoothed community policing incidents over the day show significant fluctuations with a general upward trend towards the end of the observed period. The ACF and PACF indicate some degree of autocorrelation in the data, especially at lower lags. The smoothed disorder incidents over the hour exhibit volatility with frequent spikes, reflecting the unpredictable nature of disorder incidents. The ACF and PACF calculations reveal a moderate level of autocorrelation, indicating that past incidents have a lingering impact on current incidents. The smoothed traffic incidents over the day show a stable pattern with some fluctuations. The ACF and PACF plots suggest minimal autocorrelation, implying that traffic incidents occur somewhat independently from one another. The smoothed vehicle stop incidents over the day display variability similar to traffic incidents but with more pronounced peaks and troughs. The ACF and PACF calculations highlight notable autocorrelation at lower lags.

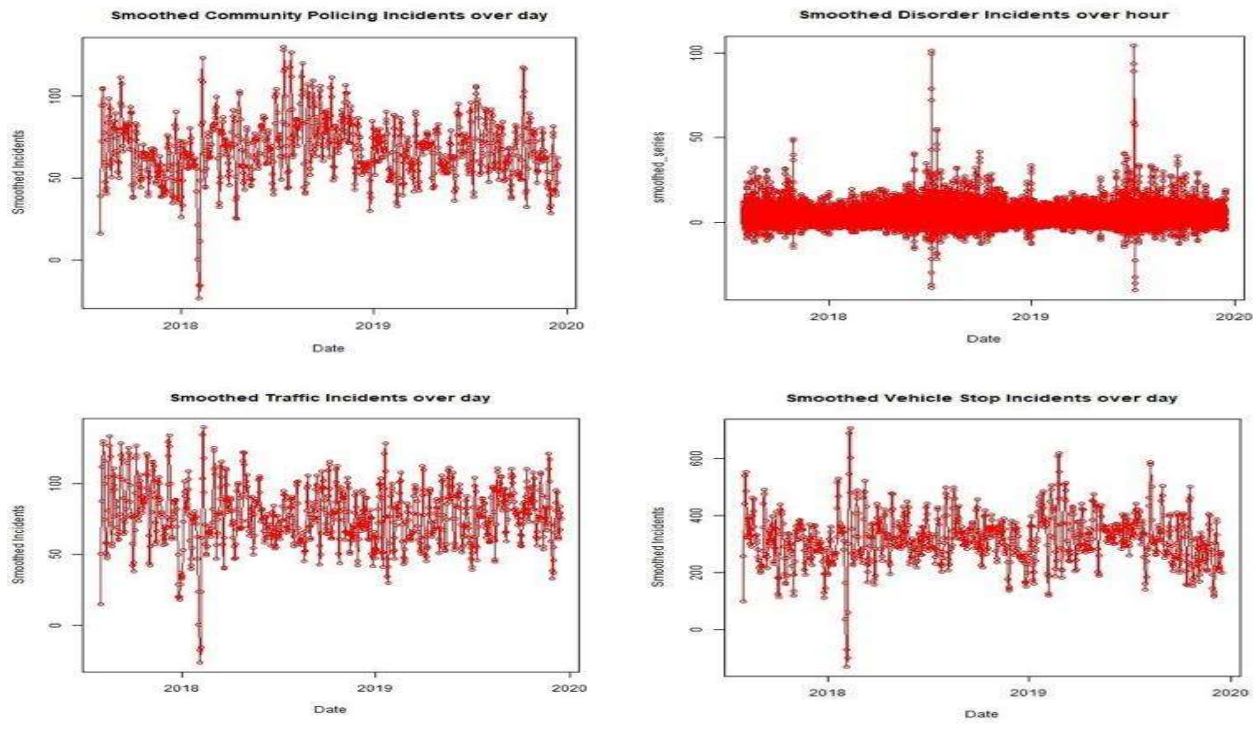

Figure 10 **- Hourly analysis of community policing, disorder, traffic, and vehicle stop incidents.**

[5] Prophet | Forecasting at scale. (facebook.github.io)

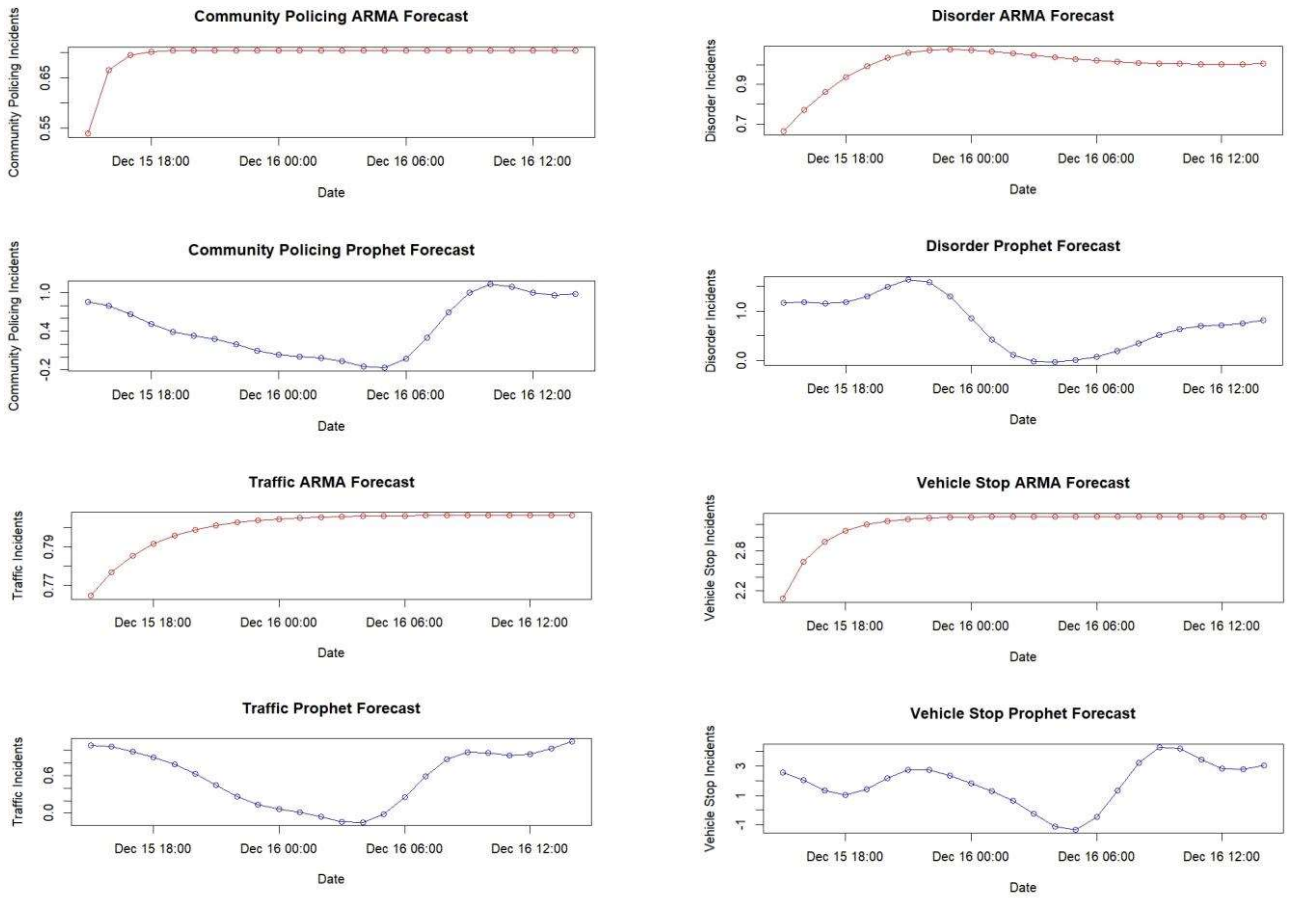

Figure 11 - Forecasting community policing, disorder, traffic, and vehicle stop incidents.

Summary

Our analysis of the Santa Clara Crime Dataset reveals that neither the monthly nor the weekly crime incident count in Santa Clara County have increased from August 2017 to December 2019. Instead, the crime incidents demonstrate daily, weekly, and seasonal fluctuations around a stable overall mean. We identified cyclic trends in eleven primary crime incident categories. We identified distinct daily and hourly patterns for six types of crime incidents. Traffic incidents, vehicle stops, alarms, and pedestrian stops are predictable through ARMA and SARIMA models with minimal autocorrelation, while disorder and community policing incidents show higher variability and multiperiodicity and require more complex modeling methods than those we had available for this project. Our ARMA and SARIMA model forecasts and diagnostic analyses suggest that while 14 models that we developed fit well, others, particularly for more volatile incident types, could benefit from refinement. Our hourly analysis highlights the unique temporal behaviors of each incident type, with community policing and disorder incidents showing significant autocorrelation and periodic trends, whereas traffic incidents occur more independently.