**Term Project Report**

**Predictive Analysis of Crime Rates in Los Angeles**

**Dhruti Joshi**

The project aims to develop a predictive model for crime rates in Los Angeles to enhance public safety measures and resource allocation. Given the dynamic nature of crime patterns, precise forecasting can help law enforcement agencies to preemptively address potential hotspots.

**Objective**: The primary objective is to accurately forecast crime occurrences across Los Angeles using historical crime data.

Data Source: - <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>

**Data Preparation**

* **Preprocessing Steps**: Data cleaning involved handling missing values, removing duplicates, and converting data types. Date and time entries were standardized, and geographical data were verified for accuracy. Features like day of the week, month, and hour of the day were engineered from date-time information.

Exploratory Data Analysis (EDA)

1. Top 10 Crime Distribution

Most of the crimes was on theft, burglary and violent offence and vehicle stolen being the most.

A colorful pie chart with text

Description automatically generated

1. **Top Areas with higher and lower crime rate.**

I have found that top 5 area which has higher crime rates are central, 77th Street, pacific, southwest, and Hollywood.

A graph of blue and white bars

Description automatically generated

1. **Heatmap of crime location**

A map of a city

Description automatically generated

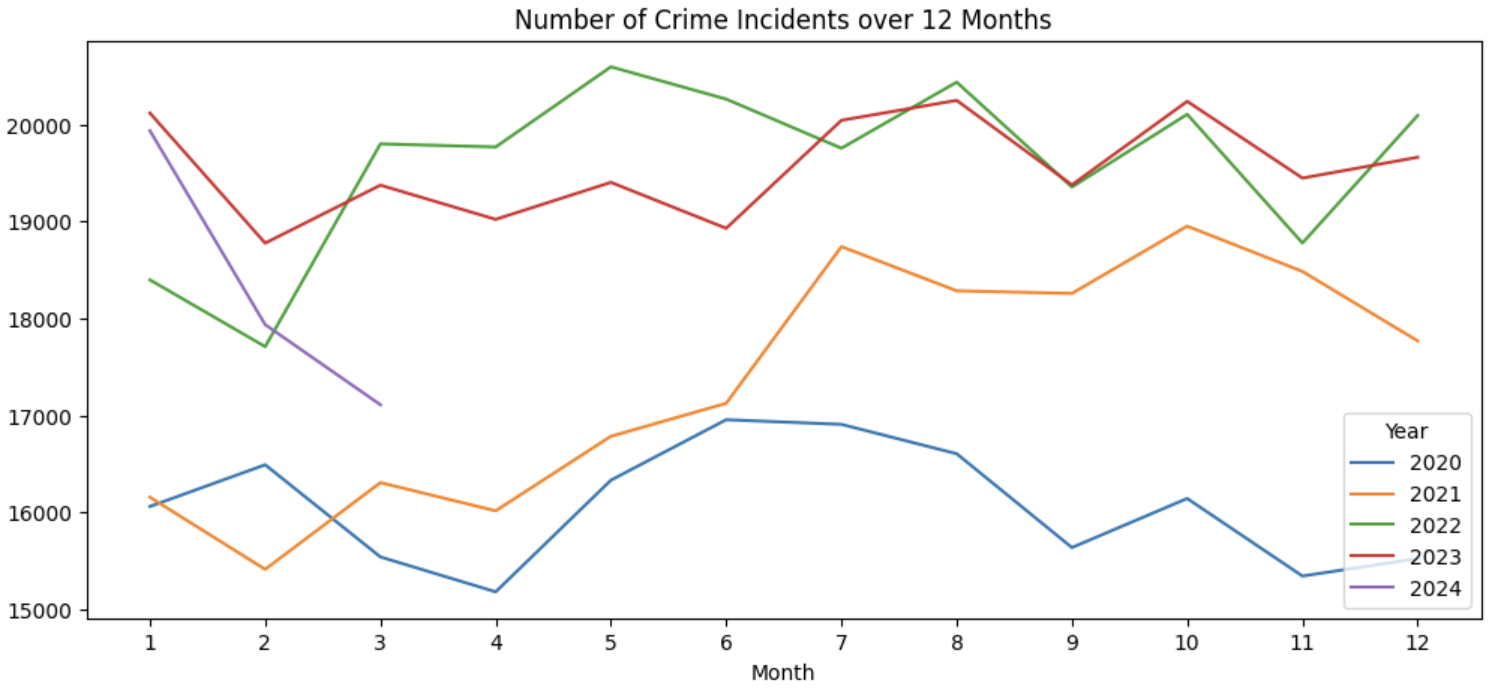
1. **Number of Crime Incidents w.r.t Victims gender.**

A graph of crime

Description automatically generated

Not much difference between the victims’ gender and can’t say anything on unknown gender so, can’t say which gender is most affected.

1. **Number of crime incidents over 12 months**



1. **Is there any pattern of the number of crime incidents over the hour**?

Yes, the number of crime incidents is the highest at noon (the lunch time). The second peak is around the dinner time.

A graph of a number of crime

Description automatically generated

1. **Analysis of 'Age':**

It shows the number of the top five descents over ages. There are two exceptional peaks: age 50 for White and age 35 for Unknown. (More explanations are waiting for experts who know more about crimes)

A graph of crime victims

Description automatically generated

A graph of crime

Description automatically generated

As for the numbers of different genders over ages, two observations:

There is a separating line at age 35. Below this age, the number of crime incidents against female is higher than that of male. There are two exceptional peaks for male: age 35 and 50. There is a gender code 'H' (82 crime incidents), which is not mentioned in the official website.

1. **Distribution of top 3 Victim Descents by Crime Area.**

A graph of a number of victims

Description automatically generated

Looking at the top 3 vulnerable demographics in LA, Latinx, White and Black communities has most suffered in the highest crime occurring areas of Central, 77th Street and Pacific. This could also be that there are lesser White people living in these two areas. White victims are highest in the 77th street, Pacific areas. Again, this could be because these areas have higher white population. Similarly, areas such as 77th street, Mission, Newton, Rampart, Foothill and Hollenbeck have a higher number of Latinx victims than the other descents.

**Autocorrelation**

The Autocorrelation function (ACF) plot or correlogram and the Partial Autocorrelation Function (PACF) plots to get a sense of which lags are significant. Sofrom the ACF plot we understand that at a 95% confidence (shown by the shaded area), only the first few lags or up to 9 lags have a significant correlation, i.e. that the series is strongly correlated only for the first couple of lags and decays after that. This could be evidence that the series is random walk We also see that there is presence of both trend and seasonality in the data.

A graph of a graph

Description automatically generated with medium confidence

**Trend**

Before models can be run it is important to check the data for stationary assumption. This is because in certain models, we don't want our data to be dependent on time. One way to check on trend is to do a seasonal decomposition on the data, especially for data that fluctuates in time.

The first subplot on the very top shows the plot for the original data with no decomposition. The second subplot shows a clear smooth trend pattern in the data. This is clear evidence of a non-constant mean.

The third subplot shows the decomposed seasonality pattern in the data. The last subplot shows the noise or residual component in the time series data.

A graph of different colored lines

Description automatically generated

I have also looked at the rolling statistics to confirm any instances of non-stationarity.

A graph of a line graph

Description automatically generated with medium confidence

From the figures above we can see that the data violates the stationarity assumption of constant mean, constant variance and constant covariance.

**Model selection**

Choosing the SARIMAX (Seasonal Autoregressive Integrated Moving Average with exogenous variables) model for time series forecasting typically revolves around specific characteristics and requirements of the data and analysis goals. Here are the primary reasons why SARIMAX I selected:

* 1. **Seasonal Patterns:** If the data exhibits clear seasonal fluctuations, SARIMAX is advantageous because it explicitly models this seasonality. The model can account for patterns that repeat over a set period, such as weekly, monthly, or yearly, by incorporating seasonal differencing and seasonal autoregressive and moving average components.
  2. **Non- Stationarity Data:** SARIMAX integrates differencing into the modeling process, making it suitable for handling non-stationary data. Non-stationarity is common in real-world time series data due to trends and seasonal effects. Differencing helps stabilize the mean of the time series by removing changes at the level of the data and seasonal variations.

The implementation involved building several SARIMAX models with varying parameters, refined through grid search techniques to optimize performance. Below is the summary of my model.

A screenshot of a computer

Description automatically generated

The SARIMAX model results summarize the fit for a seasonal ARIMA model with the configuration (1, 0, 2)x(2, 0, [1], 12) applied to a crime count dataset with 31 observations spanning from January 2020 to July 2022. The model's AIC and BIC values are 520.602 and 530.640, respectively, suggesting the relative quality of the model fit given the number of parameters used. The coefficients for the AR and MA parts at different lags indicate how past values and errors influence the model predictions. Notably, the P-values for many of the seasonal components are very high, indicating they may not be statistically significant and could potentially be simplified or removed. The diagnostics, including a Ljung-Box test with a p-value of 0.29 and a Jarque-Bera test with a p-value of 0.59, suggest that the residuals are fairly random, and the model does not suffer from major issues like autocorrelation or non-normality, although the effectiveness of the seasonal components should be reevaluated.

**Plot Diagnostics**

The residuals of a model are the resulting error terms from the difference between the observed value and the predicted or fitted values. By looking at the residuals we can check whether a model has satisfactorily captured the information in the data. For a good forecasting method, the residuals will pass the following assumption tests of:

1. Uncorrelated residuals. -- tells us that the model captures all information in the data
2. Residuals with zero mean. -- tells us that the model forecasts are not biased

A collage of graphs

Description automatically generated

The plot diagnostics is the several ways of aspects of the residuals from a SARIMAX model fit to ensure it is appropriately modeling the data. It generates diagnostic plots that include the analysis of standardized residuals, histogram plus estimated density for normality, normal Q-Q plots, and the autocorrelation function (ACF). These diagnostics help verify that the residuals do not exhibit patterns or autocorrelation, are normally distributed, and conform to the assumptions necessary for reliable forecasting.

**Below is the plot the predictions against the true observed values.**

A graph with a line and a red circle

Description automatically generated

Looking at the plot we can see that the predicted values follow the observed values very closely but not mimicking the trend and seasonality in the time series which says that models still need to improve.

**Performance Analysis**

**Evaluation Metrics**: The models were evaluated using Root Mean Squared Error (RMSE) to quantify the prediction accuracy. The final RMSE was significantly lower than the baseline models, indicating an improvement in predictive accuracy. The Root Mean Squared Error of the forecasts is: 656.30.

The Root Mean Squared Error (RMSE) of 656 indicates that while our model performs adequately, there is considerable room for improvement. So, in next steps we can gather more data which can help to make a stronger prediction and we can use incoming data to evaluate the model’s forecasting performance.