**Crime Data Analysis-Los Angeles**

Regis University

MSDS696\_X70\_Data Science Practicum II

Professor John Koenig

By

Sangita Lamichhane

2022

**Project Overview**

The funding trend of increasing police spending was reversed for the first time in years by reducing the Los Angeles Police Department (LAPD) budget by $150 million in 2020. Utilizing the data on crime made available by the [City of Los Angeles](https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z)from January 2010 - June 2019, and from 2020 to present we used univariate time series modeling to forecast reported crime in Los Angeles to determine recent crime trends in Los Angeles.

This project analyses the available crime data by simple exploratory data analysis and using Machine learning to predict the crimes that occur in Los Angeles. Developing the models and analyzing with AR, MA, ARMA, and SARIMA models we determined the best fit with the lowest AIC score.

**Graphical user interface, chart

Description automatically generated**

Figure Crime Trend during the time period of study

**Research Problem**

Los Angeles has one of the largest police departments in the country, with jurisdiction over almost 4 million people. Through this project analysis we aim to identify which areas and populations in LA have been hit the hardest by crime in past few years, so that agencies can reallocate budget resources to social services in those high-impact communities. Furthermore, this project studied the following from the data:

* Crime changes during the time of study
* Areas in LA with highest reported crime.
* Top impacted demographics in LA

**Data Description**

The crime data [Los Angles Crime data](https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z) for this project is made available by the city of Los Angeles. The data reflects incidents of crime in the City of Los Angeles from 2010 to present. There are 2,714,774 observations over 28 variables. Each row provides a record of information on crime that was reported, location it occurred, time it occurred, crime type, and information on the victims of the crime.

**Methodology**

Through the exploratory data analysis, we identified the top demographics, areas, and crime changes during the period of study. Additionally, the analysis was done using a univariate time series problem that helped in predicting the monthly average crime count for each year. Time series model iterations included AR, ARMA and SARIMA algorithms. We compared the models against each other using the AIC score and evaluated the final model with RMSE. An AIC score determines which machine learning model is best for a given data. The RMSE simply measures how accurate predictions are. We predicted data through 2019 to guide decisions for the business context.

**Final Model Performance**

For the model analysis, along with our time series data analysis for stationary we also performed additional analysis in stationarity in the residuals of the time series. Then, by performing a seasonal deposition we identified the trend and seasonality in the data.

Chart

Description automatically generated

Figure Crime forecast Diagram

Working with AR, MA, ARMA, and SARIMA models we determined the best fit with the lowest AIC score. Then evaluated our best model, a SARIMA model, using the best AIC. We have taken the AIC score 722.89 and RMSE of 51.8.

**Table

Description automatically generated**

Figure ML models with their respective AICs

**Conclusion**

From our explorative data analysis, our model predicted a slight decrease in crime rate from 2017 onwards. The funding for LAPD should be redistributed to social service solutions to public safety problems in the high crime areas such as 77th Street, Southwest, and Pacific.

**Chart, bar chart

Description automatically generated**

Figure Crime distribution in LA over several locations

We can see the top vulnerable communities in LA is Black and Latina communities. The highest crime area seems to be 77th street and Southwest in LA. White community seems to be least vulnerable but one of the reasons could be there are less white population compared to others in those area.

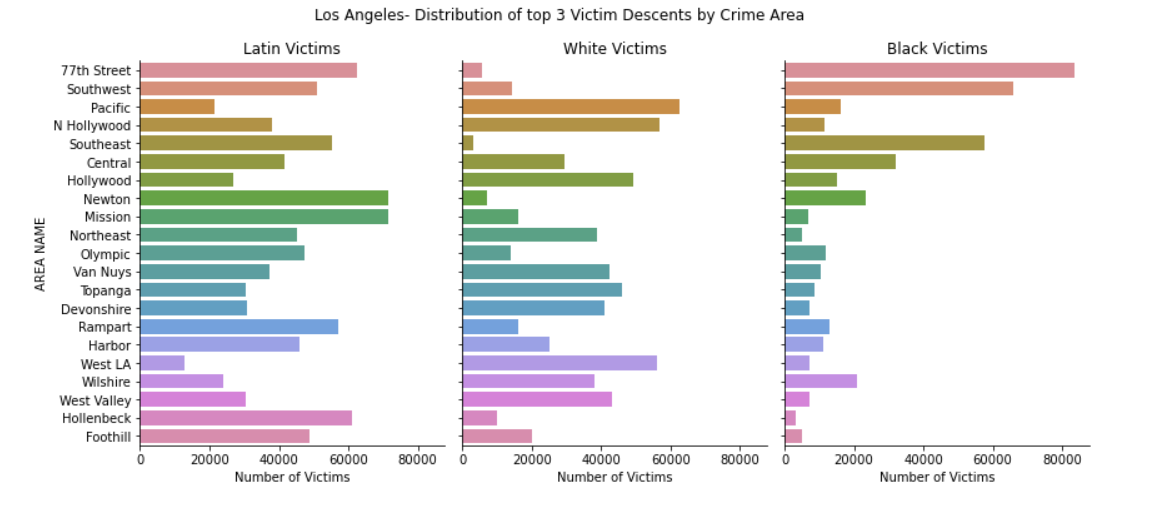
****

Figure Crime distribution in LA over different races

Year 2017 saw the highest average number of crimes reported, while 2013 saw the lowest average number of crimes reported in LA. Additionally, we can see decreasing trend of average crime from 2017 onwards.

**Chart, box and whisker chart

Description automatically generated**

Figure Average Monthly Crimes Reported per Year in LA

**Future research and Recommendation**

Based on this analysis we recommend gathering additional historical data which could help to make stronger predictions and forecasts further into the future. Furthermore, including exogenous variables from census data to add complexity to the univariate model will be insightful. We can also implement a Facebook Prophet model for comparison and model evaluation. Additionally, incoming data could also be used to evaluate our model's forecasting performance.

**References**

Ademos. (2022, August). *Chapter 23: Using ARIMA for Time Series Analysis.* Retrieved from ademos: <https://ademos.people.uic.edu/Chapter23.html>

Data Source: <https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z>

<https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>

David Bruce, S. K. (2021). *Los Angeles Crime Forecasting*. Retrieved from LA crime forecasting: https://github.com/sidneykung/LA\_crime\_forecasting

Prabhakaran, S. (2021, August 22). *machine learning+*. Retrieved from ARIMA Model – Complete Guide to Time Series Forecasting in Python: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/