Los Angeles, California Crime Data (2010-2019)

Emanuel Calderon, Jiayuan Ai, Jonathan Huck, Monika Malik, and Zalan Mumtaz

Table of Contents

Executive Summary…………………………………………………………………………... 3

Problem Statement……………………………………………………………………………. 3-5

Methodology………………………………………………………………………………….. 5

Sample,Explore and Modify…...…………………………………………………..… 5-7

Model……………………………………………………………………………….... 7

Assess…………………….………………………………………………………….. 7-8

Results……...………………………………………………………………………………... 8-10

Conclusions and Recommendations……………………………………………………….... 10-11

References…………………………………………………………………………………... 12

Appendix……………………………………………………………………………………. 13-17

**Executive Summary**

Crime in America is the second leading cause of death among youths. Los Angeles has the third highest crime rate in the nation directly affecting four million youth in the city. The LA Police precinct requires swift responses to crime issues & reasons for these occurrences. Through the analysis of a large number of criminal record data that occurred in LA from 2010 to 2019, we obtained statistical analysis of variables such as crime location, time, and criminal information. We found the correlation between these variables and then established a prediction model. Specifically, the LAPD will be able to both predict if crimes were created by a Juvenile or an adult and create an understanding of correlations between crime categories. Additionally, the Los Angeles police can understand which areas have a higher crime rate, and they can further predict the crime situation in 2020 through the model. With the information, the police can formulate their next working strategy to monitor the streets of the community efficiently to ensure the safety of society, and to reduce the occurrences of crime. At the same time, through model prediction, it can reduce unnecessary consumption of resources and maximum efficiency through limited resources. Resources including installing more equipment and increasing the number of patrolling police in key areas of concern.

**Problem Statement**

The Los Angeles Police Department (LAPD) consists of 10,002 officers as of 2019. It is the third-largest municipal police department in the United States with 18 divisions and 21 police stations throughout the city. Between 2010-2019 there have been about 2.11 million crimes reported to the LAPD. According to Neighborhood Scout, a search engine that creates neighborhood profiles, the crime rate in LA is substantially higher than the national average across communities in America. Relative to California, Los Angeles has a crime rate that is higher than 86% of the state’s cities and towns of all sizes. With numbers like this, residents worry about the impact this will leave on their children and community.

Economically, high crime rates affect communities in many different ways. The price of properties is affected because homeowners find it difficult to sell their property. The Center of American Progress reported that a 10% reduction in homicides would lead to a 0.83% increase in housing values the following year. Additionally, home insurance premiums can rise because of vandalism and theft. Aside from housing, crime impacts local businesses because of repeated thefts and loss of income due to the costs of repairing damage from vandalism.

High crime rates have a great impact on the youth of the community. The United States Census Bureau reported that Los Angeles has a population of approximately four million with 21% are under 18 and in danger of falling into the wrong hands. Violence is the second leading cause of death for young people in America and the leading cause of death for young people in every major city. With crime being in their neighborhood they will look up to the “leaders” of the neighborhood which most likely be drug dealers, thugs, and pimps. Youth violence increases the risk for behavioral and mental health difficulties, including future violence perpetration and victimization, smoking/substance use, academic difficulties, and school dropout.

Our team decided to analyze the LAPD’s reported crime dataset that contained all reported crimes from 2010 to 2019. With this information, we want to classify if a crime can be committed by Juveniles for unsolved and new case analysis, and proactive intervention to reduce youth crimes. We want to know what areas need an increase in patrols, surveillance, and equipment to keep the streets of Los Angeles safe, keep the officers safe and prevent the downfall of communities and the future of our youth.

**Methodology (SEMMA)**

When first exploring the data, there were upfront changes that needed to be addressed in order to make sound predictions. The data had around two million rows spanning from dates 2019 to 2010 and 28 variables (variables seen in Appendix, Figure 1). The idea was to obtain predictability of the given information to model and predict the path to help the officers of the LAPD in 2020.

**Sample, Explore and, Modify**

The team remained in the sampling, exploring and modifying phase for quite some time due to various dimensions in the data. As seen in Appendix (Figure 2), we initially started our sampling and exploring phase (in the left, about 2 million rows) and ended with the data of about 297,000 rows (in the right). Note, although there are many variables seen, only 7-9 variables were actually used in the classification modeling phase. Additionally, we created a summary datasets to analyze clusters and predict time series.

The team immediately realized there were many redundant variables and items that needed transformation for the analysis. Variables such as date reported & date occurred were transformed to date, month and year. Also, variables such as Crime Code, Mocodes, Premis Codes, Weapons used codes had too many categories for any reliable predictive analytics. Lat/Lon was helpful for Geographical plots. The team realized we may have more columns than required for modeling due to the fine comb detail provided by the police department.

Appendix (Figure 3) shows that even after removing redundant variables, it was still difficult to discover anticipated and unanticipated relationships. It seemed as though there was still a lot of room to resample the data and look at it in other angles for reliable prediction power.

When attempting to cluster, partition or even create principle components to simplify given inputs, the team ran into many difficulties with computing power, time and incorrect variable type names as inputs. The team had to spend more time resampling and exploring the data to make better sense of what the data could help to explain. It was at this point the team decided to bucket the data rows to reduce the detailed information provided but provide us about six to seven new columns of useful patterns. The crime description, premise description, weapon description, and status description received new variables to eliminate the complexity of the data and provide the team with a better trend. The team also expanded on the time, day, month and year of the occurrence of the crime. After bucketing the variables we were able to see better trends and correlations in the data. Next, we needed an approach to reduce the volume of data.

In manufacturing and in the six sigma process, one would collect data for a project to attempt to find the bad performance of processes around the plant. When performing the data search of the triggers to the bad defects, one would go as far back as five years of the life of a plant since any further look in the history might steer the project leader in the wrong direction due to changes performed in machines, leadership, and other unknown variables. It was this that led the team into the decision of narrowing down how far back the data used should look. The team decided to look back as far as five years (down to 2014) from 2019 and then continue to predict the 2020 data. This decision has led the team with faster computing time and more use of other predictive modeling techniques. The next version of the data set was reduced to one million rows.

Additionally, the team decided to take all the variables and summarize the data. The team evaluated the slimed data with the predictions of the cluster model and to evaluate trends in a time-based series. The clustered data is seen in Appendix (Figure 4). JMP allows the user to join the grouped 21 precinct locations with the new crime, weapon and premise categories in order to assist in fully understanding the trends being seen. This produced over 20 extra columns and eliminated the complexity of the one million rows by bunching the data up and relating them to the areas (the initial intention of what the group was aiming to assist the officers with). This helped greatly reduce the complexity of the original given data and gave the team a chance to predict an outcome for 2020. Due to this summarization, the team was able to spot trends to assist with later unsupervised modeling (see Appendix Figure 5).

The last step to move the team to create supervised models was the application of the validation columns. With the assistance of reducing the variables/rows to obtain meaningful data, it was the use of the status description that separated out the three fields. Using the juvenile 1’s and zeros, the team used the stratified function in the creation of the validation column to separate out the training, validation, and test. The stratified sampling technique was used as the number of 1’s in the Juvenile column was only ~4% of the total data (Appendix Figure 6).

**Model**

Jump Predictor Screening utility was used to analyze the contribution of 9 variables selected as part of the initial analysis. Looking at the ranking and contrition of these variables, the team was able to drop Month and Day to be used in our classification problem. The final variables used for classification Models were - Premise, Crime, Victim Age, Victim Sex, Weapon and Hour of Occurrence. Six models namely - Generalized Regression, Neural Network, Decision Tree, Boosted Forest, Boosted Tree and Naive Bayer’s Classifier were created to predict if the criminal was a Juvenile or not.

For the Time Series Model, the team created Seasonal ARMA, Moving Average and Winter Model. Seasonal ARMA was selected as the final model based on the accuracy of the results. Results of the model are available in Appendix (Figure 7).

For district clustering, Ward, Centroid and Average hierarchical models were created. Ward algorithm represented the most meaningful clusters. Based on the results, 11 districts were categorized as Violent Crime HotSpots. Results of the model and list of crime hotspot districts are available in Appendix (Figure 9).

**Assess**

When assessing the classification models the team compared every model's value of true positives from the test partition in the confusion matrix. The goal was to find the model with the highest value of true positive from the test partition because it will tell us what model has the most accurate prediction. Since all misclassification cannot be equally considered, our team will be focusing on the accuracy of the prediction of true positive for each model. We also considered the area under the ROC curve because if the value breaks away from one, the model will start to make inaccurate predictions. In theory, we want the area under the ROC curve to be as close to one as possible. Next, we analyzed which model is not overfitting or force-fitting. This needs to be considered because if an overfitting model were to be chosen, it would not be able to work as well as expected with completely new data. We were also interested in minimizing the expected cost of classification error, but we did not have any associated statistics to compare the cost of misclassification. This should be considered in future analysis. Lastly, we tried to compare the complexity and accuracy of the models to see if we need to make any trade-off.

**Results**

Historical crime data is considered to be highly unpredictable as the team found this out through modeling. After thoughtful consideration of crime classification through supervised and unsupervised learning, the data set illustrates a few main findings.

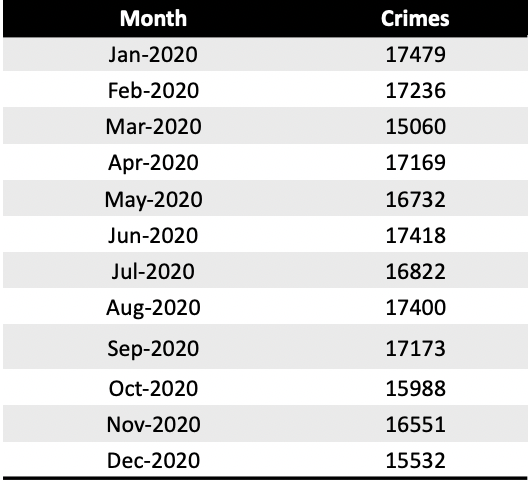
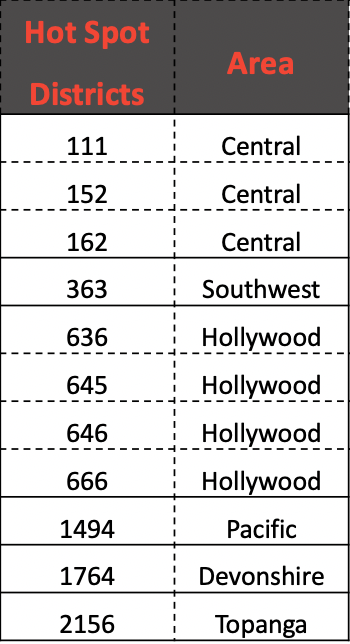
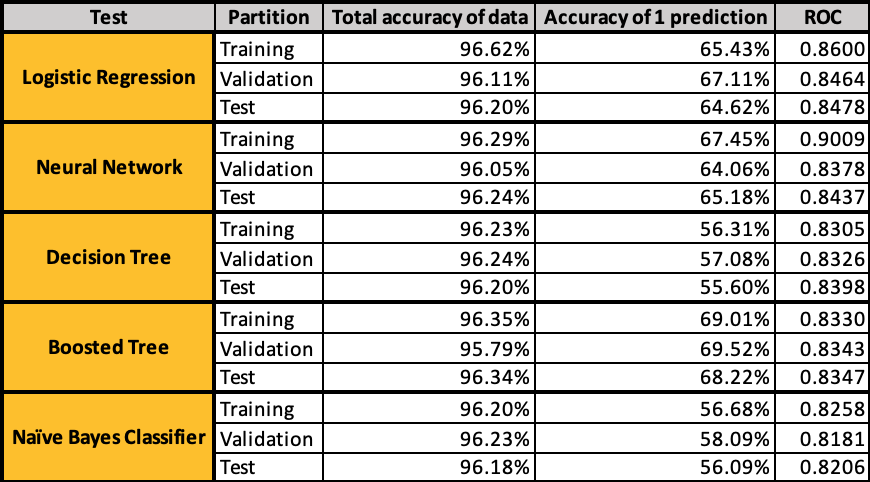
 

Table above shows the predicted number of crimes in LA in 2020. These numbers are based on historical trends and seasonal variations. Because of the current lockdown situation in LA because of COVID-19, these might not be very reliable. This is one of the issues with time series modeling of Crimes as there are various external factors which affect the actual statistics and are difficult to incorporate in any model. The Clustering model resulted in precisely outlining violent crime hot spot districts. These areas are to be carefully considered for further examination by the precinct to combat the crimes.



Finally, our main predictive models for Classification - Logistic Regression, Neural Network, Decision Tree, Boosted Tree & Naive Bayers Classifier demonstrated measurable lift in predicting if a crime is committed by a Juvenile. We concluded the Boosted Forest model prediction may be lacking some validity because of added complexity. The Boosted Tree Model with sixteen trees, proved to have the best predictive power at 68.22 percent accuracy of a one prediction and an RMSE of .1779. Additionally, it did not show any overfitting problem as seen in other Models. If we needed to make a tradeoff between complexity and accuracy because of the volume of data, we could have selected Generalized regression.

**Conclusions and Recommendations**

The classification model should be presented to the LAPD to help with unsolved crimes and enlist a comparative study of the new approach to target youth crimes.

Through statistics and analysis of crime information in 2010-2019, you can find Data shows violent crime hot spots. We found that the most common causes of crime are drugs, blunt objects and deadly objects. And there is a certain correlation between these reasons. The prediction of the crime rate in 2020 shows that the highest crime rate is expected to be in January 2020, while the lowest predicted rate is in March 2020. The lowest crime rate in 2020 is expected in March. We have also discovered hotspot crime areas, such as the 77th area has the most occurrences, and, in 2015-2019, crime increased in the central region. The Time series model confirmed a crime drop from the end of 2017 to now.

For the analysis results, our suggestion is: we can predict the number of crimes in specific areas within 0-6 months through time series, which will greatly help the police to make plans, they can better plan how to in different areas Assigned equipment and personnel. We can also train neural networks and compare their results with the existing data so that we can confirm the accuracy of the prediction and modify it in time. Secondly, we can also model the time by the hour, or divide the crime data in a year by season. Different time points and seasons may have special reasons for the crime category and crime rate to change. Make predictions more accurate. At the same time, we can also model housing data, which can explain more contributions to crime, such as what types of houses are prone to theft and which areas are prone to robbery, which can bring us more thinking directions. Finally, the improvement of local weapons laws can also bring more convenience to police law enforcement, because many criminal incidents are accompanied by the use of guns.

**References**

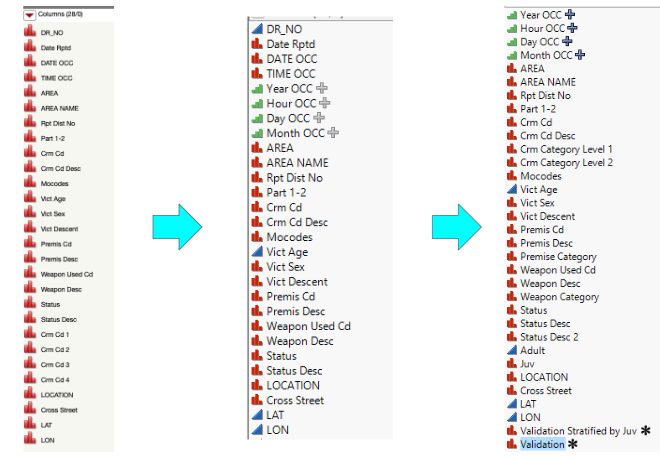
1. Los Angeles, CA Crime Rates. (n.d.). Retrieved April 10, 2020, from <https://www.neighborhoodscout.com/ca/los-angeles/crime>
2. Maximino, M. (2017, February 16). The impact of crime on property values: Research roundup. Retrieved April 10, 2020, from <https://journalistsresource.org/studies/economics/real-estate/the-impact-of-crime-on-property-val>[ues-research-roundup/](https://journalistsresource.org/studies/economics/real-estate/the-impact-of-crime-on-property-values-research-roundup/)
3. The Los Angeles Police Department. (n.d.). Retrieved April 10, 2020, from <http://www.lapdonline.org/>
4. U.S. Census Bureau QuickFacts: Los Angeles city, California. (n.d.). Retrieved April 10, 2020, from <https://www.census.gov/quickfacts/fact/table/losangelescitycalifornia/RHI125218>

**Appendix**

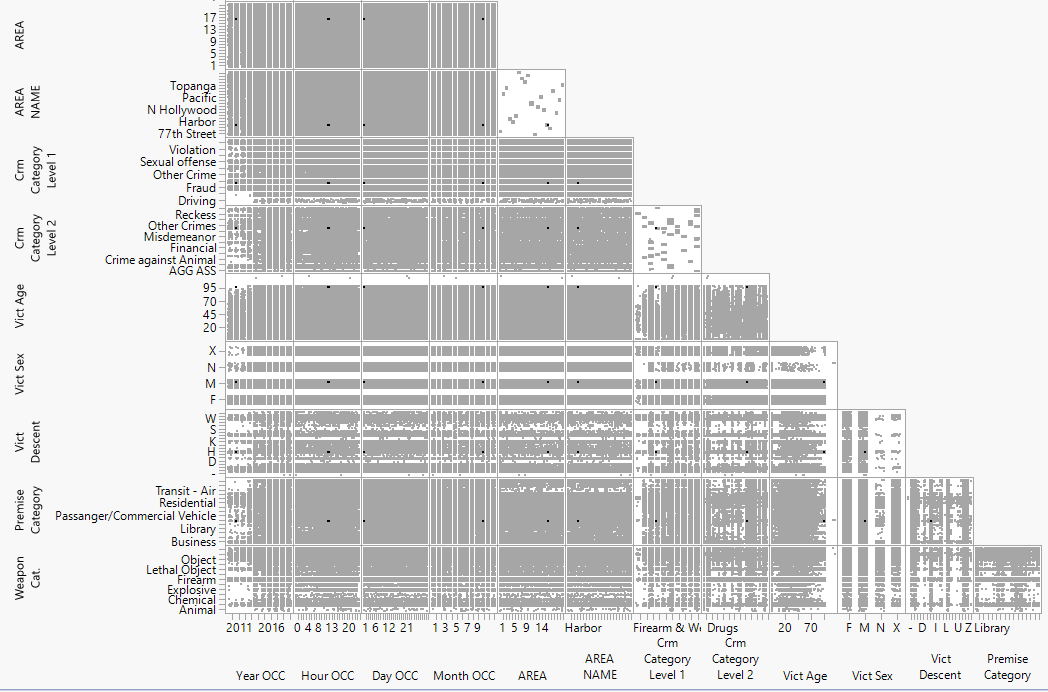
**Figure 1- Columns in the initial Data Set**

****

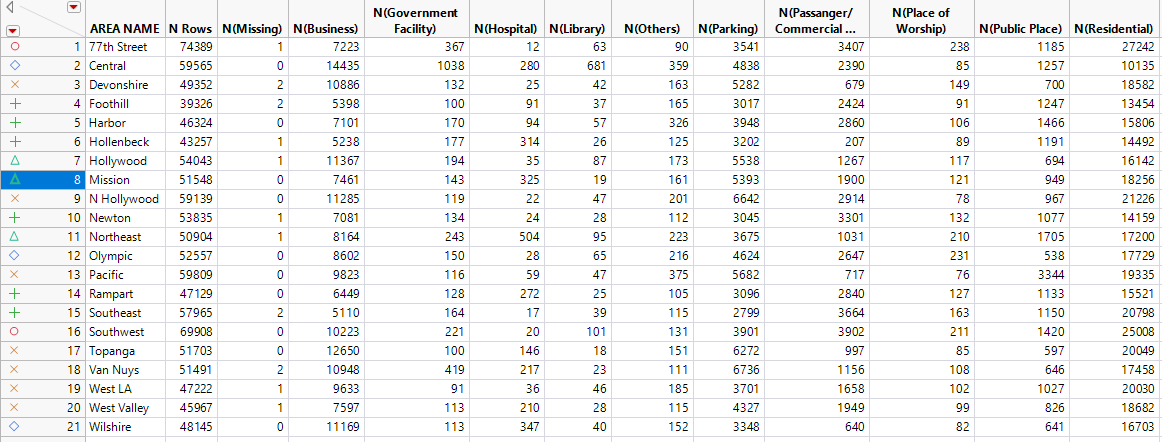
**Figure 2 - Column Transformation**

****

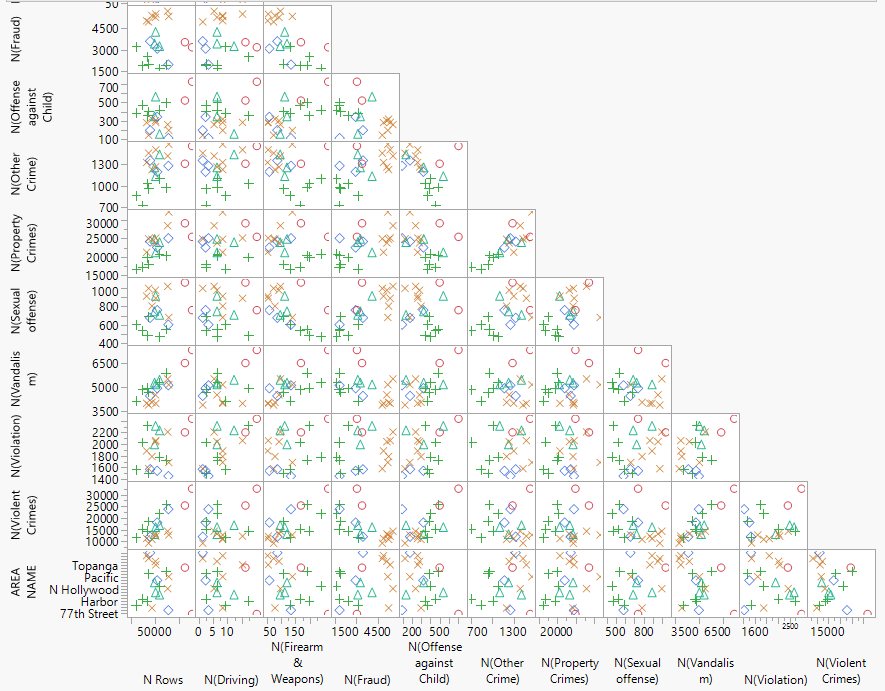
**Figure 3 - Scatter Plot matrix of predictors**

****

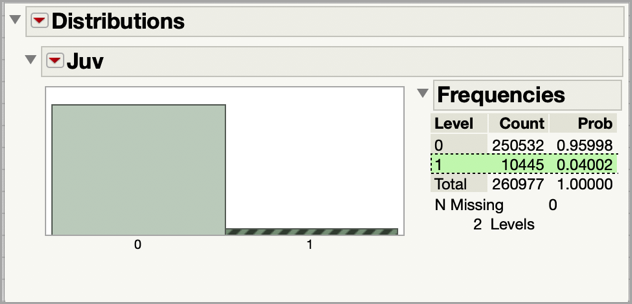
**Figure 4 - Resampled data set, grouped area name and subgroup variables**



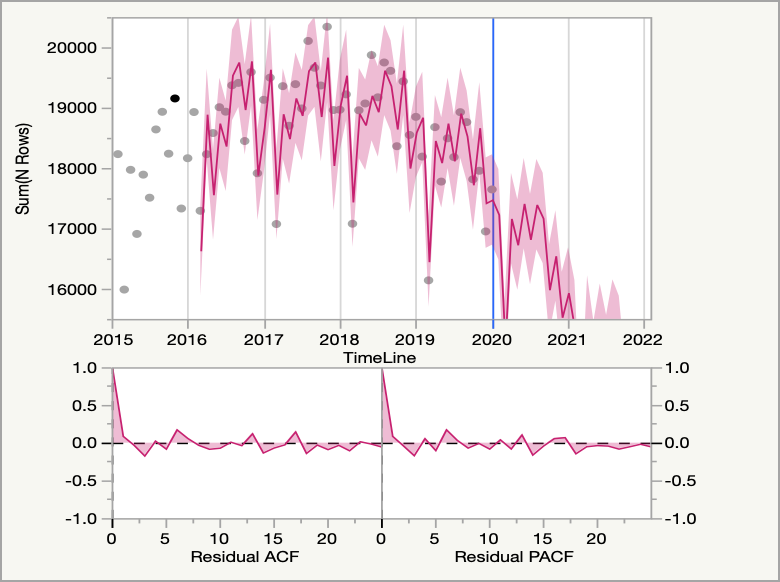
**Figure 5 - Scatterplot matrix tables summarized data set**

****

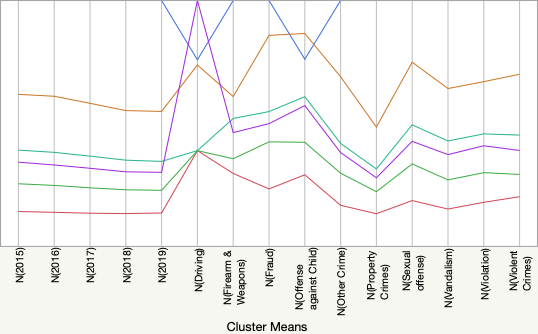
**Figure 6 - Distribution of Juvenile Crimes**

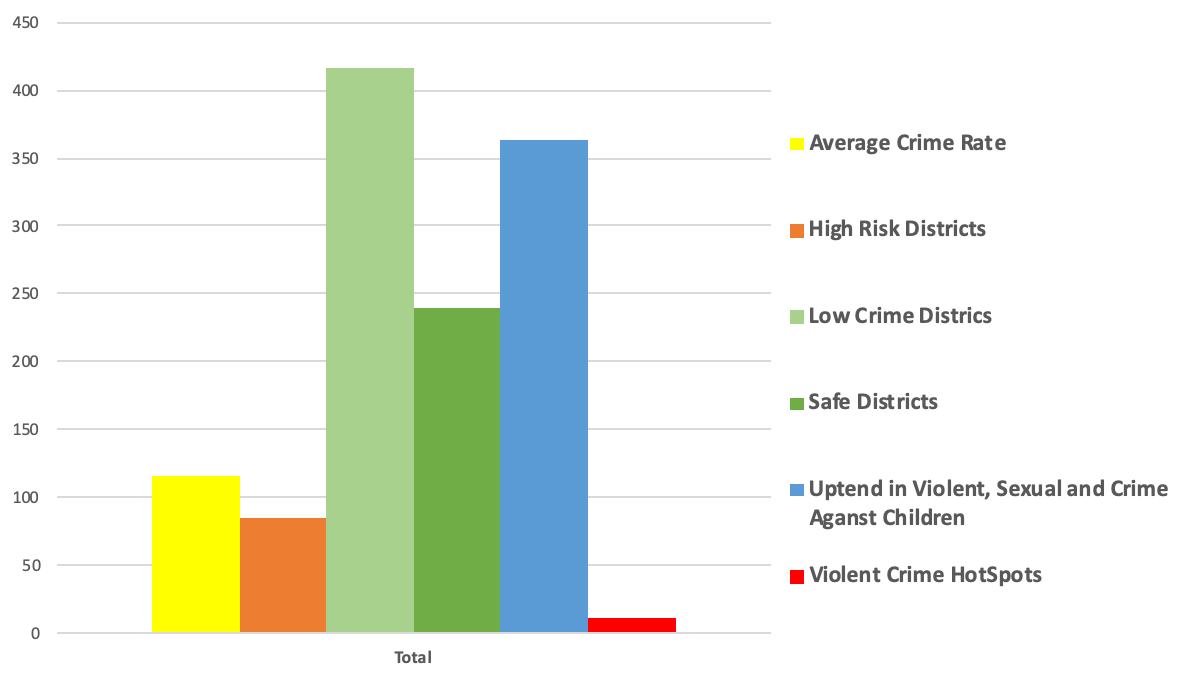


**Figure 7 - ARMA Season Time Series Model**

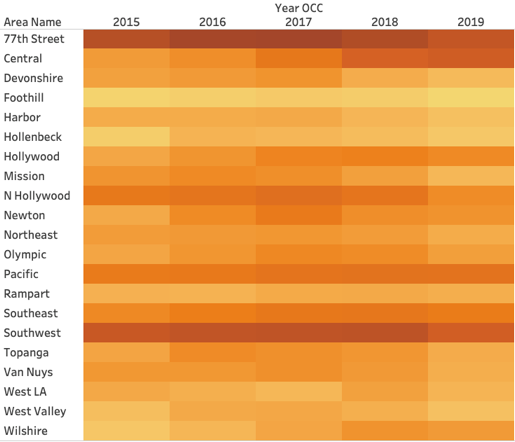


**Figure 8 - Hierarchical Ward Clustering by Reporting District**

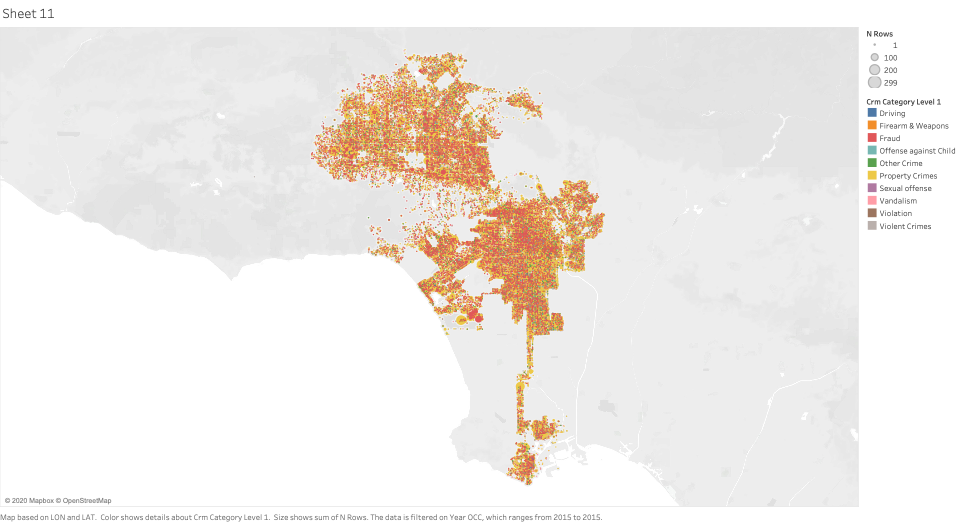




**Figure 9 - Crime by Area and Year**

****

**Figure 10 - Geographical Profile of Crimes**

****