Crime in Los Angeles

Madison Razook
University of Colorado Boulder
mara5374@colorado.edu

David Stone
University of Colorado Boulder
david.stone-1@colorado.edu

1 Abstract

Using the City of Los Angeles' "Crime Data from 2010 to the Present", this paper sought to find interesting patterns in the victimhood of violent and property crimes. The Los Angeles Times, which maps crime rates in Los Angeles County, defines violent crime as homicide, rape, aggravated assault, and robbery.[1] Property crimes include burglary, theft, grand theft auto, and theft from vehicle.[2] Specifically, we applied association rule mining and random forest classification to find the relationships between the type of crime committed and victim characteristics such as age range, sex, race, and location. We found that sex and race play a large role in predicting whether or not a crime was violent, with Black and Latina women being especially at risk for violent offenses. These findings can be used for work in the areas of victim advocacy, crime matching, and law enforcement.

2 Introduction

Los Angeles is the US's second most populous city, with LA County being the most populous county in the country. The area is diverse in terms of both ethnicity and income.[3] Examining what happens in this city will not only benefit its millions of residents, but may also shed some light on crime trends in other areas of the United States. Through our analysis of this data, we sought to find the relationship between victim age, sex, ethnicity and the type of crime they were targeted for. The answers to this research question may help create a short-term solution for crime, in the form of safety advice for those individuals who are most at risk. To truly address crime at its core, however, we must also look at the perpetrator. Understanding who is targeted and for what creates a clearer picture of the offender's motivations.

3 Related Work

The Los Angeles Police Department (LAPD) devotes a section of its website to research.[4]There are several reports that detail crime rates for 2013-2015, list possible reasons for these numbers, and explain the LAPD's efforts to combat crime. Additionally, the 2016 year-end crime statistics subsection contains charts that document the number and type of crimes committed during the last three months of the year in each division of the city, along with their percent change between months. The last items in the LAPD's research section are year-end use-of-force reviews for 2015 and 2016. The information here examines how the LAPD's treatment of suspects varies according to suspect race, affiliations, and other categories. While the information here is not as extensive as the eight-year data set we will use for our project, it does highlight the most recent crime trends for us. Knowing LAPD's recent tactics may also help us understand any patterns in the arrest rates that we discover.

The Los Angeles Times has also used data maintained by the City of Los Angeles to create public sources of information. In 2009, they launched their Mapping LA project, which provides maps and statistics for all of LA County's neighborhoods.[5] Of interest to our work is the subproject Crime LA.[6] Crime LA is a map showing violent and property crime rates for each neighborhood, highlighting those areas which have recently seen an increase in either type of offense. As of this writing, data from the February 20th-February 26th, 2018 time period provided the map's most recent update. In addition to this service, Crime LA also contains rankings for each neighborhood based on the number of violent or property crimes per 10,000 people, taken from the July 31, 2017- January 28, 2018 time period. Similar to the LAPD's research, this resource quickly updates readers on the most recent criminal activity, but is not able to show long-term crime trends.

A 2010 data analysis done by the Los Angeles County Department of Public Health looked at the effects of homicide on life expectancy by neighborhood and ethnicity in LA County.[7] The study used cause elimination techniques on mortality records and population estimates. Results indicated that the South Service Planning Area (SPA) of LA had higher poverty levels and percentages of black and Latino residents compared to other areas. Homicide was estimated to reduce the life expectancy of black males by 2.1 in LA County and by almost 5 years in low-income urban areas. 82.4% of homicide deaths victimized people between 15-44 years of age, with the majority of these deaths caused by a firearm. The study concluded with several strategies for homicide reduction based on this information, noting that high homicide rates are often correlated with low levels of social cohesion. With more recent data, we can see if young black males are still most at risk for homicide deaths, and in which areas they are particularly vulnerable. These results could be used to measure the effectiveness of any homicide reduction strategies used since 2010.

A 2013 study published in the University of Pennsylvania Law Review measured the effects of zoning on crime in LA.[8] The authors first looked at crime rates in eight neighborhoods with high levels of crime but different forms of zoned land use. A second method compared two groups of neighborhoods with similar crime trajectories, with one group experiencing zoning changes while the other did not. Zoning changes, mostly in which parcels were converted to land use, led to a significant reduction in crime. The study concluded that mixing residential-only zoning with commercial blocks may serve to reduce crime overall. The results of our project may suggest crime trends that corroborate or disprove this idea.

A 2015 research paper published by the University of Colorado at Boulder took a similar focus to our work [9]. The team analyzed crime data from both Denver and Los Angeles using apriori, Naive Bayes classification, and decision tree classification, while providing an evaluation of each method. Their research purpose, however, was to identify spatiotemporal hotspots for crime in both cities. Our work used similar methods to identify common characteristics of victims.

Keyvanpour et al did not use the same crime data as this paper, but rather provided an overview of crime data mining techniques [10]. They found that using the neural networking technique of self-organizing maps can extract features from large datasets filled with text values. These feature maps can then be used for efficient k-means clustering and subsequently in crime matching. Crime matching is the process of matching potential suspects to unsolved crimes through data mining the circumstances of the crime. Since the characteristics of the victim are a major component of the crime, we hope that analysis in that area can further aid crime matching efforts.

4 Data Set

Our dataset consists of reported and documented crimes with the LAPD dating back to 2010. It includes the date of the crime, the date it was reported, the victim's gender and descent, area where the crime occurred, among other attributes. https://catalog.data.gov/dataset/crime-data-from-2010-to-present

5 Main Techniques Applied

All work for this project was done in Jupyter Notebook, with code from the libraries Pandas, NumPy, Matplotlib, mlxtend, and scikit-learn.

5.1 Preprocessing

We first had to do significant cleaning and preprocessing of our data before applying any mining techniques to it. To begin with, we selected only a subset of the attributes: "DR Number" (the case number), "Date Occurred", "Time Occurred", "Area ID", "Crime Code", "Victim Age", "Victim Sex", "Victim Descent." Other attributes were either redundant, e.g. "Location" and "Address" were not

Figure 2: Histogram of Victim Sex values

needed when we already had "Area ID", or irrelevant to our research purpose, such as "Weapon Used Code." Once we had isolated our relevant attributes, we began data transformation and reduction. values"Date Occurred" first had to be converted to a Pandas datetime object. To more closely analyze the circumstances of each crime, we split the "Date Occurred" values into three new columns: "Year Occurred", "Month Occurred", and "Day of Week."

Each of these new attributes had their values converted to strings. Since “Month Occurred” had 12 unique values, which could hinder frequent pattern mining, we grouped its values into “Q1”, “Q2”, Q3”, and “Q4.”

“Time Occurred” had many unique values, all originally formatted as integers in military time, e.g. “2300.” Again, to increase our chances of uncovering frequent patterns, we reduced the large number of values by grouping them into 4-hour intervals.

“Area ID” is an integer with a range of 1-21. The LAPD divides the city into 21 areas, each belonging to one of 4 bureaus: Central, South, West, and Valley. We mapped each Area ID integer to its corresponding bureau.

“Crime Code” is an integer code for the type of crime. We looked up what each code stood for and re-labeled the values as either “Violent” or “Property”.

“Victim Age” was reduced to groups of 10 years each (e.g., 21 is mapped to group “Age 20-29”). “Victim Descent” consisted of one-letter codes for each ethnicity which we grouped these into their larger racial classifications; for example, “C” (“Chinese”) was mapped to “Asian.” “Victim Sex” had values “M” and “F”, with Unknown/Null values indicated by “H”, “X”, or “-”. We cleared out the latter values, leaving us with only “Male” and “Female” for our “Victim Sex” category.

We continued our data preprocessing by dropping all null values. At the end of this phase, with the exception of “DR Number” and “Date Occurred”, our dataset had only string values. During earlier attempts at data preprocessing, we mapped our values to integers. However, we eventually settled on methods in which string labels were preferred.

Finally, using Matplotlib, we generated some histograms to get a better understanding of the trends in our data.

The graphs show that at least among reported crimes, Hispanic/Latinx and Age 20-29 are the most common characteristics for victims, with an even split for Male and Female.

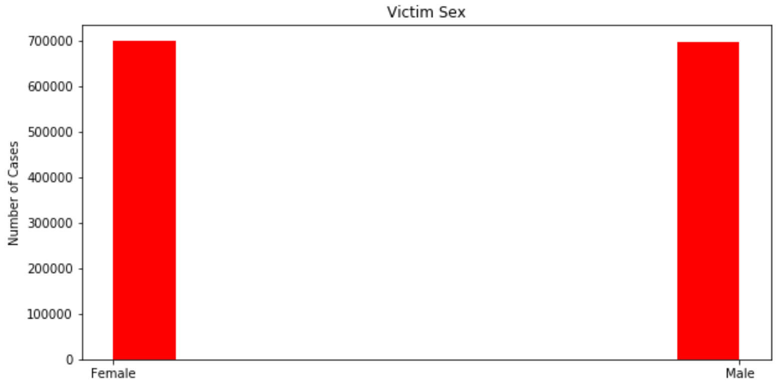


Figure 1: Histogram of Victim Sex

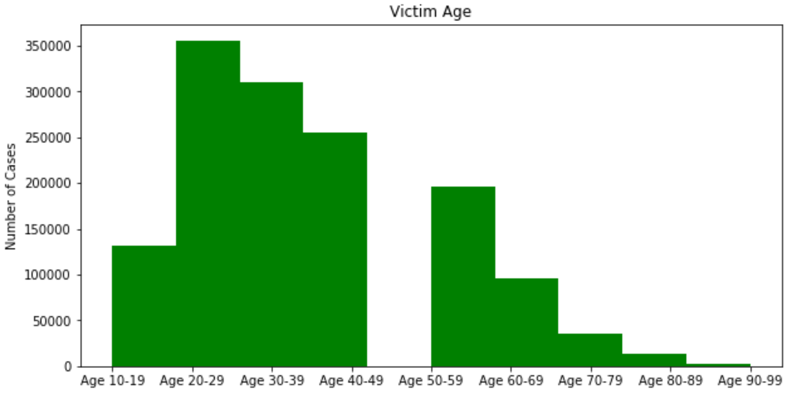


Figure 2: Histogram of Victim Age

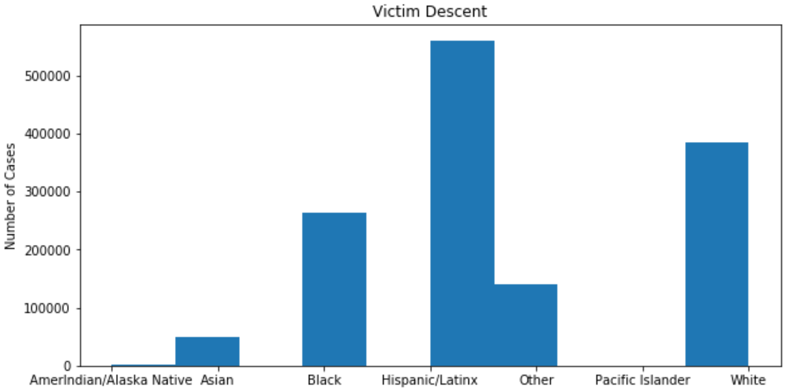


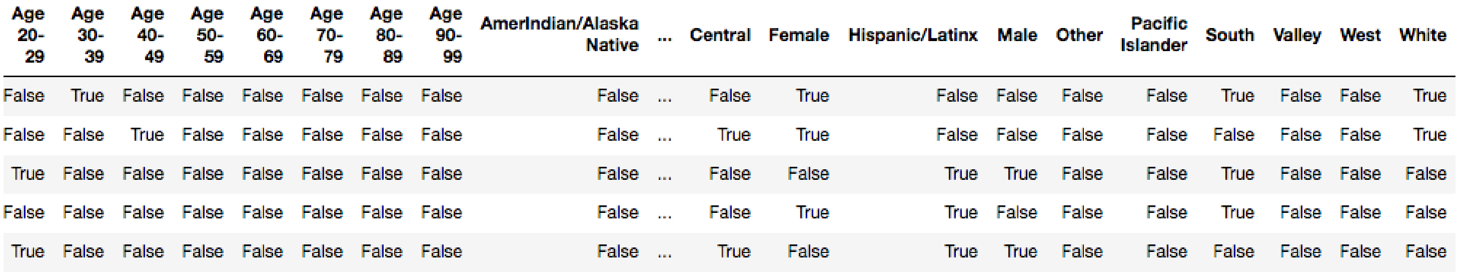
Figure 3: Histogram of Victim Descent

5.2 Association Rule Mining

To determine which demographics are at most risk for violent and/or property crime, we mined for frequent patterns using the popular apriori algorithm. For this phase, we used the open-source Python library mlxtend. First, we separated our dataset into cases of violent crime v_data and cases of property crime p_data. Victim information “Victim Age”, “Victim Sex”, “Victim Descent”, and “Area ID” were compiled into a vector for both of these new datasets, labeled as v_victim and p_victim. We did the same for the temporal attributes “Year Occurred”, “Month Occurred”, “Day of Week”, and “Time Occurred, but these did not yield any frequent patterns.

Using the apriori algorithm required us to further transform our data. Since the algorithm is most commonly used to (but is not limited to) analyze transaction data, we changed our datasets to follow a transaction-style format in which columns are each unique value of each attribute, and values are “True” or “False.”

Figure 4: Example of transaction encoding on v_victim data.

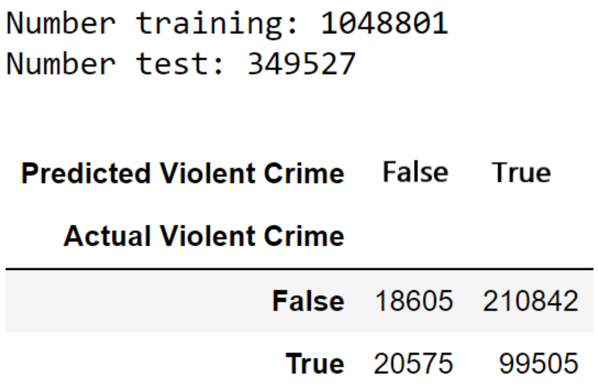


Age 20-29	Age 30-39	Age 40-49	Age 50-59	Age 60-69	Age 70-79	Age 80-89	Age 90-99	AmerIndian/Alaska Native	...	Central	Female	Hispanic/Latinx	Male	Other	Pacific Islander	South	Valley	West	White
False	True	False	False	False	False	False	False	False	...	False	True	False	False	False	False	True	False	False	True
False	False	True	False	False	False	False	False	False	...	True	True	False	False	False	False	False	False	False	True
True	False	False	False	False	False	False	False	False	...	False	False	True	True	False	False	True	False	False	False
False	False	False	False	False	False	False	False	False	...	False	True	True	False	False	False	True	False	False	False
True	False	False	False	False	False	False	False	False	...	True	False	True	True	False	False	False	False	False	False

Once this had been completed for v_victim and p_victim, we ran the apriori algorithm to find frequent patterns with a minimum support of 10%. From these patterns, we generated association rules, and subsequently filtered out any rules that did not satisfy the conditions lift > 1, confidence > 0.5, and leverage > 0.01. We chose these conditions to ensure that each itemset found was positively correlated. These rules are discussed in the Results section.

5.3 Classification

As discussed in the association rule making, we mapped out our data and made every variable into a boolean true or false statement. Instead of displaying the date and time as one time object, we separated each day of the week, quarter of year, and five hour time interval which all consisted of a true or false value. We did this with all variables which allowed us the option to make a decision tree for classification. Specifically, we used the Random Forest Classifier. The Random Forest Classifier is an Ensemble algorithm that uses two or more algorithms together to classify objects. This specific Ensemble algorithm creates a set of decision trees from a randomly selected subset of the training set and then aggregates the votes from those decision trees and decides the the final class of the test object [11]. To implement Random Forest, we need to pick the dependent variable or simply what we would like to predict. We chose to predict whether or not the crime committed would be a violent crime or a property crime based off of many attributes from our data. The attributes that gave us the best results were Victim Sex, Victim Descent, Area, and the Month the crime occured. By splitting on these 4 attributes, we correctly predicted whether or not a crime was a violent crime with an accuracy of almost 70%. This data is shown as a confusion matrix in figure 5 below.



```
Number training: 1048801
Number test: 349527
```

Predicted Violent Crime	False	True
Actual Violent Crime		
False	18605	210842
True	20575	99505

Figure 5: Confusion graph showing accuracy of classifier

Unfortunately, after much trial and error, the above attributes yielded the best accuracy for predicting violent crime. Furthermore, when predicting other variables such as the victim's gender or the time of year based off of the remaining attributes, we could not get an accuracy that we deemed significant or numerically over 50%. Future work could be done to explore all different classifiers to yield significant results.

5.4 Clustering

In earlier stages of our project, we used k-means clustering for the following figures. This approach did not yield the clear results we were looking for, but did help us decide where we should focus our research and analysis.

Our thought process for figure 6 was looking for a significant difference in the victim's gender and the type of weapon used in the crime. 100s are for guns, 200s for knives, 300s for blunt objects, 400s for fists and feet, and 500s for miscellaneous weapons. The dataset had four options for gender: male, female, and two non-binary categories H and X. Each option was assigned a numeric value: 0 for male, 1 for female, 2 for H, and 3 for X. The results seem to say that the weapon type does not depend on the gender of the victim.

Figure 6

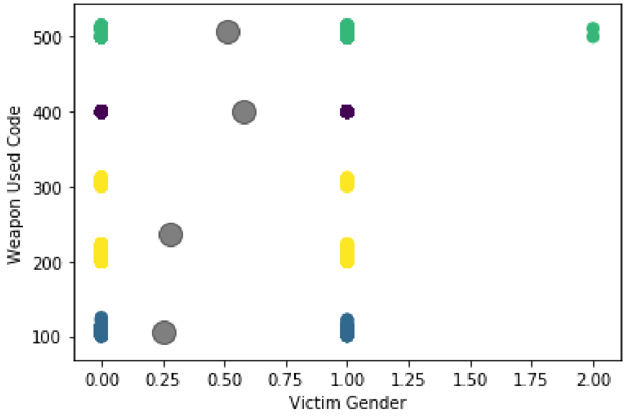


Figure 7 shows the comparison of the victim's age and where the crime occurred. Area ID 1 is for Central L.A. and is the most densely populated and has the highest amount of crime out of around 20 neighborhoods within the city limits. The results do not point towards different neighborhoods having higher crime towards a certain age group.

Figure 7

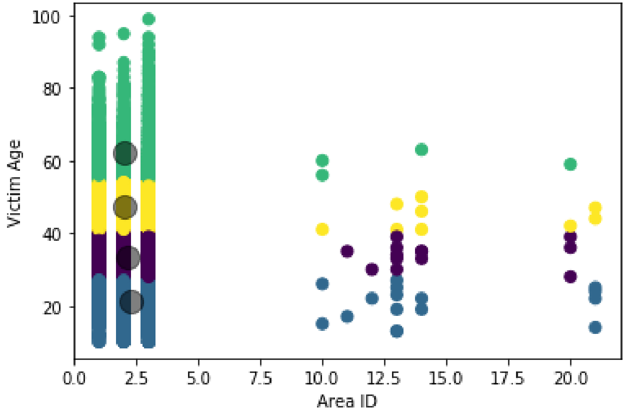
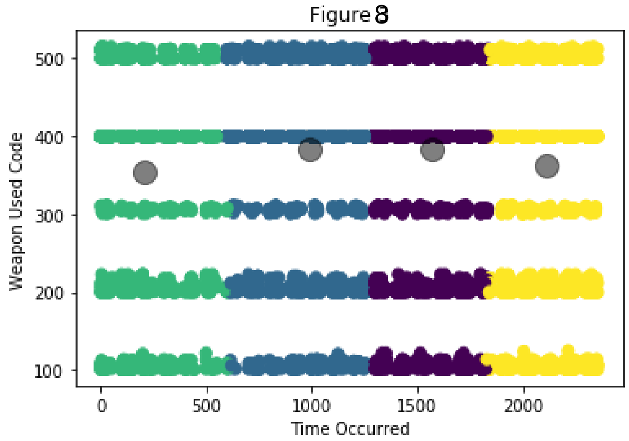
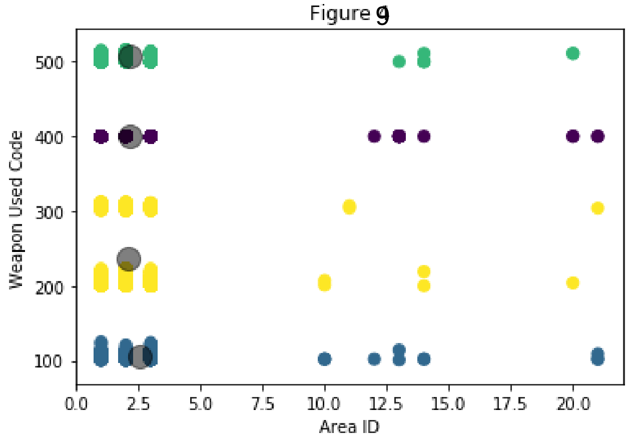


Figure 8 shows the comparison of the type of weapon used and the time the crime occured. This was trying to answer the question of whether or not different weapons were used at different times. The graph shows there is no time of day one weapon is used more than the other.



With Figure 9, we attempted to find an interesting connection between area and weapon used. Areas in the lower range tended to see more weapon use, with 400s being the least used category; beyond that, the clusters here are consistently sized and offer little information.



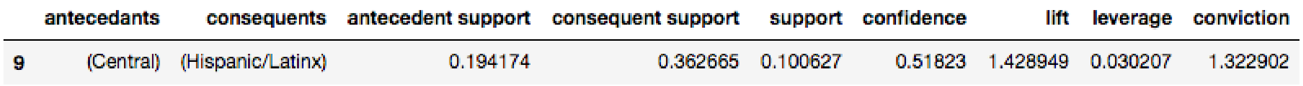
6 Key Results Found

Figure 10 Association Rules for Property Crime:

Our association rule mining found far more rules for violent crime than it did for property crime. For victims of property crime, the only rule fitting our conditions was {Central -> Hispanic/Latinx}, indicating that this demographic is especially at risk to be targeted for property crime. It is very important though to note that after fully preprocessing our data there were over 900,000 reported property crimes and just over 480,000 violent crimes. Our violent victimhood rules found a strong correlation between living the South area of LA and being Black. {Central -> Hispanic/Latinx} makes an appearance here as well, although its lift is not as high as {South -> Black}. Continuing down the list, we begin to see more of a pattern with {South -> Female}, {Black -> Female}, and {Age 20-29 -> Female}. These rules together suggest that Black women in their 20s living the South area of LA are highly at risk to be a target of violence. Additionally, it is worth noting that {Female} appears more frequently in the violent crimes dataset than any other characteristic, with a support of almost 59%. It is followed by {Hispanic/Latinx} which has a support of 47%.

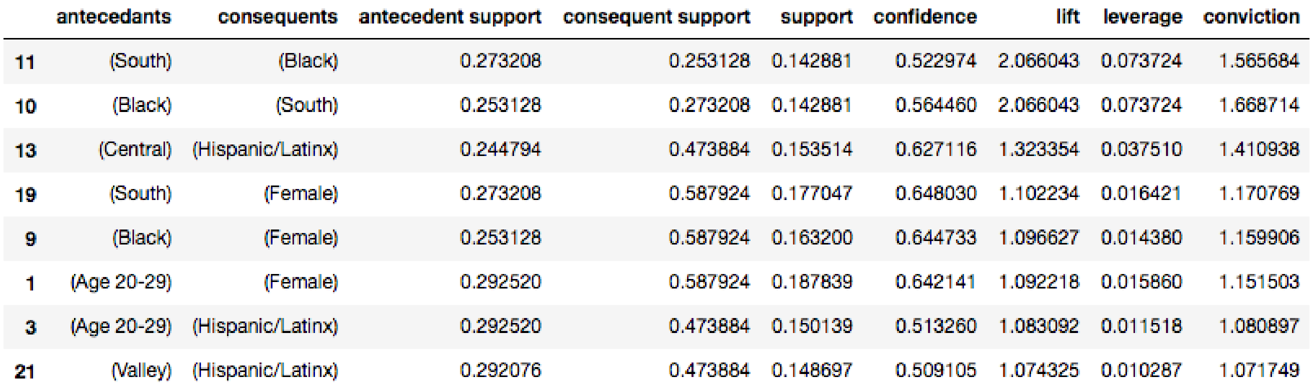
When classifying, we also found that the Random Forest classifier highlighted race and gender as its most important features in classifying. Association

example, with the confusion graph in figure 5, although Victim Descent, Area ID, Victim Sex, and the Month the crime occurred all helped the classifier with its prediction, when calling the most important features function, Victim Descent and Victim Sex accounted for more than 65% of the influence while the Area ID and the month the crime occurred in only had the remaining 35% of influence. It is important to note the frequency of only two genders as well as the majority of descents being White, Black, or Hispanic/Latinx compared to the many Area ID's and Months where the crimes occurred and could be attributed to why the algorithm weighted Sex and Descent more than other attributes.



	antecedants	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
9	(Central)	(Hispanic/Latinx)	0.194174	0.362665	0.100627	0.51823	1.428949	0.030207	1.322902

Figure 11 Association Rules for Violent Crime:



	antecedants	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
11	(South)	(Black)	0.273208	0.253128	0.142881	0.522974	2.066043	0.073724	1.565684
10	(Black)	(South)	0.253128	0.273208	0.142881	0.564460	2.066043	0.073724	1.668714
13	(Central)	(Hispanic/Latinx)	0.244794	0.473884	0.153514	0.627116	1.323354	0.037510	1.410938
19	(South)	(Female)	0.273208	0.587924	0.177047	0.648030	1.102234	0.016421	1.170769
9	(Black)	(Female)	0.253128	0.587924	0.163200	0.644733	1.096627	0.014380	1.159906
1	(Age 20-29)	(Female)	0.292520	0.587924	0.187839	0.642141	1.092218	0.015860	1.151503
3	(Age 20-29)	(Hispanic/Latinx)	0.292520	0.473884	0.150139	0.513260	1.083092	0.011518	1.080897
21	(Valley)	(Hispanic/Latinx)	0.292076	0.473884	0.148697	0.509105	1.074325	0.010287	1.071749

7 Applications

Identifying the most at-risk groups for violent and property crimes in LA is the first step for more concrete action. In particular, victim advocacy groups should keep in mind that young Black and Latina women in the South, Central, and Valley area of LA are more likely to need their services than LA women of other demographics. Police departments could also use this knowledge when patrolling areas of LA. Since the victim's identity comprises an important component of a criminal's MO, this analysis could also be used in crime matching efforts. Additionally, our paper only made a distinction between violent and property crime which focused on the victims of these crimes and not the perpetrator of these crimes and we chose this to give a more specific idea of crime in Los Angeles. With twice as much property crime as violent crime or two thirds of the overall crime in the last decade, our research can be used by media outlets and the general public to eradicate general bias of crime and create a well informed public of the exact types of crime happening where they reside. Further research is needed to find patterns between victim identity and specific crimes such as domestic violence or vandalism and whether or not arrests were made for each crime.

REFERENCES

[1] Los Angeles Times. 2018. Violent Crime Ranking - Mapping L.A. Retrieved from http://maps.latimes.com/neighborhoods/violent-crime/neighborhood/list/

[2] Los Angeles Times. 2018. Violent Crime Ranking - Mapping L.A. Retrieved from http://maps.latimes.com/neighborhoods/property-crime/neighborhood/list/

[3] U.S. Census Bureau. 2018 Quick Facts: Los Angeles City, California. Retrieved from https://www.census.gov/quickfacts/fact/table/losangelescitycalifornia/PST045216

[4] Los Angeles Police Department. 2018. Research. Retrieved from http://www.lapdonline.org/research

[5] Los Angeles Times. 2018. Mapping L.A. Retrieved from http://maps.latimes.com/neighborhoods/

[6] Los Angeles Times. 2018. L.A. Crime Maps - Mapping L.A. Retrieved from http://maps.latimes.com/crime/

[7] Redelings, M., Lieb, L. & Sorvillo, F. J Urban Health (2010) 87: 670. DOI: https://doi.org/10.1007/s11524-010-9470-4

[8] Anderson, James M., John M. MacDonald, Ricky Bluthenthal, and J. Scott Ashwood. Reducing crime by shaping the built environment with zoning: an empirical study of Los Angeles. *University of Pennsylvania Law Review* 161, no. 3 (2013): 699-756. JSTOR: http://www.jstor.org/stable/23527820

[9] Tahani Almanie, Rsha Mirza, and Elizabeth Lor. 2015. Crime prediction based on crime types and using spatial and temporal criminal hotspots. *International Journal of Data Mining & Knowledge Management Process (IJDKP)* 5,4 (July 2015)

[10] Mohammed Reza Keyvanpour, Mostafa Javideh, Mohammed Reza Ebrahimi. 2011. *Procedia Computer Science* 3 (872-880)

[11] Savan Patel. 2017. Chapter 5: Random Forest Classifier. Medium.com. Retrieved April 30 2018 from https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1.