# Theft in Los Angeles
# LAPD Group

Eric Chuu Edward Fu Zhiyuan (Tyson) Ni Zelos Zhu

## I. Introduction

Video evidence of police violence and misconduct that depicts abuse of power are published everyday throughout social media and are often the topic of major news outlets. This can be detrimental to the reputation of police officers and the Los Angeles Police Department (LAPD) is no exception. The LAPD, which ranks as the third largest police department in the United States behind only New York City and Chicago, has had its fair share of controversy in the past which include the Rodney King Riots of 1992 and the Rampart Scandal which exposed widespread corruption among the Community Resources Against Street Hoodlums (CRASH) anti-gang unit. Mayor Eric Garcetti has made troves of data, which include LAPD Crime data, available to the public to increase transparency and to raise awareness among the citizens of Los Angeles. With data available from police reports, including crime descriptions and times of occurrence/report, we decided to look into the nature of some of the crimes being committed, in hopes of identifying trends that may be helpful in promoting public safety.

## II. Research Questions

Some of the research questions we sought to answer are given as follows:

1) What does the distribution of crime look like in Los Angeles? For example, is burglary more common in west LA? Is arson more likely near downtown?
2) Does time play any sort of factor in crime? For example, does car theft happen most in April?
3) Can we merge these time and geographic-based findings into some sort of predictive model? Can we predict where and when the crime will be?

## III. Data and Variable Description

Our data came from three sources:

1) LAPD Crimes and Collision data 2012-2016
2) US Census
3) Zipatlas which contained income data by ZIP code

## IV. Statistical Methods

### A. Exploring and Subsetting the Data

During the exploration stage, we noticed that our dataset included over a million observations. We decided to analyze a subset of the data, so we took the 6 most common "theft" based crimes from `Crm.Desc`, the variable with information on the crime description. This resulted with the following:

1) Shoplifting - Petty Theft ($950 & Under)
2) Theft from motor vehicle - petty ($950.01 & Over)
3) Theft Plain - Petty ($950 & Under)
4) Theft-grand ($950.01 & Over)
5) Burglary
6) Burglary from vehicle

Now we had a dataset that was about $\frac{1}{4}$ the size of the original at approximately 270,000 observations. After that, we realized that only 4 columns of the LAPD dataset were substantial: date occurred, time occurred, Crm.Desc, Location.1 (which we split into latitude and longitude). The rest of the other data was hard to utilize; we knew neither the LAPD reporting district boundaries nor how the LAPD defined "areas," so we removed these from our analysis. The status of the crime investigation was not of particular interest either. These removals left us with a need to supplement our original dataset.

### B. Crime vs. Time Exploration

Since the crime data we were given were accompanied by time and date information associated with each entry, we first did exploratory data analysis and visualization to get some sense of the temporal distribution of the crimes that we were investigating. To get some interpretive results that could be used to see any potential trends, we looked at both frequency and proportion of crime committed over the following time periods:

1) Hour of the Day, $1 \leq h \leq 24$, where $h = 24$ corresponds to midnight
2) Time Period, $T \in [1, 5]$, where $T$ is defined below
3) Month of the Year

$$T = \begin{cases} 1 & \text{if } 1 \leq h \leq 5 \\ 2 & \text{if } 6 \leq h \leq 10 \\ 3 & \text{if } 11 \leq h \leq 17 \\ 4 & \text{if } 18 \leq h \leq 21 \\ 5 & \text{if } 22 \leq h \leq 24 \end{cases}$$

In considering the frequency and proportion of crimes against each respective time period, we have visually split the bar plots to proportionally represent each of the 6 crimes. We first consider the proportion of crimes committed per hour of the day. As seen in Figure 1, it is clear that during any given hour there appear to be differences in what the dominant crime is. In the middle of the day, the proportion of `BURGLARY FROM VEHICLE` is rather small compared to that of `BURGLARY` and `THEFT PLAIN`

– PETTY crimes. However, later in the day, the proportion of BURGLARY FROM VEHICLE grows larger, at the expense of BURGLARY and THEFT PLAIN – PETTY. This should not come as a surprise, however, since crimes during the day are limited because of increased visibility and a larger civilian presence. Thus, cars are less easily stolen without drawing unwanted attention. As stores begin to close and people get off work, locked stores and occupied homes become a higher-risk target for criminals, while cars become the primary target. This shift can be seen in the barplot below.

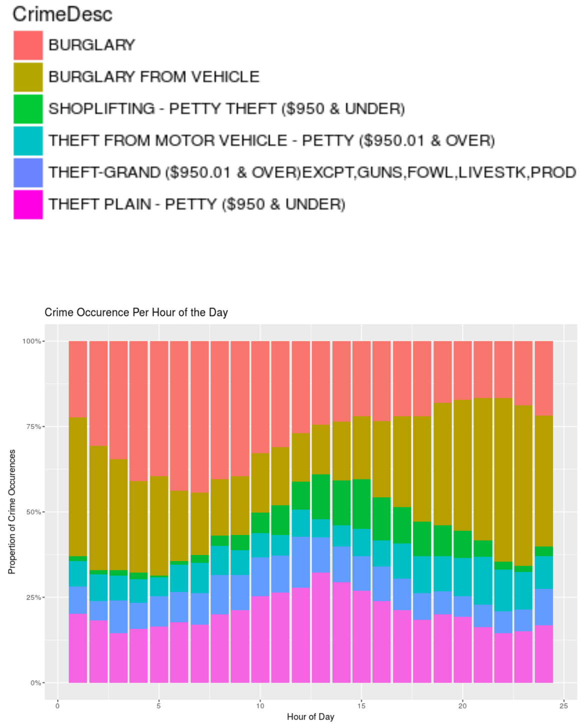

Fig. 1: *Proportion of crimes committed per hour of the day. Each bar is split into the proportion of each crime committed during that hour.*

In order to get a bigger sense of the shifts in distribution of the crimes committed throughout the day, we also divided the day into five time periods, each meant to represent common behavior that is generally uniform in peoples lives. For example, people generally spend the hours between 1 AM and 5 AM sleeping, while many people are at work between the hours between 11 AM and 5 PM. Similar generalizations were made to create the time intervals. The results can be seen in Figure 2. The trends that we witnessed before are accentuated if we consider the plot showing the proportion of crimes per time period. BURGLARY FROM VEHICLE remains highly concentrated during later hours of the day, while THEFT PLAIN – PETTY peaks in the middle of the day before reaching lower levels at night. It is also interesting to note that proportion of the other crimes, THEFT–GRAND and THEFT FROM MOTOR VEHICLE, remain relatively uniform throughout the day, which suggests some sort of inelasticity of these crimes. If we now look at just the raw frequency of these crimes over these time periods (Figure 3), it is immediately evident that crimes are predominantly committed mid-day to late afternoon. This trend is consistent with our intuition, as during these hours, many people are at work, houses are generally vacant, and stores are open. This combination, though unavoidable, inevitably results in a period of high crime.

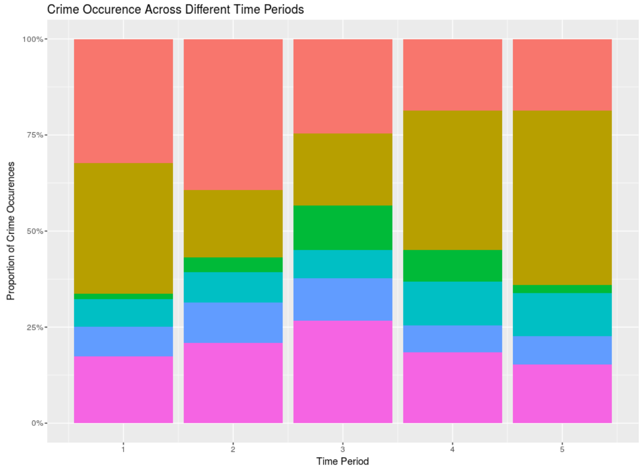

Fig. 2: *Proportion of crimes committed per time period. Each bar is split into the proportion of each crime committed during that hour.*

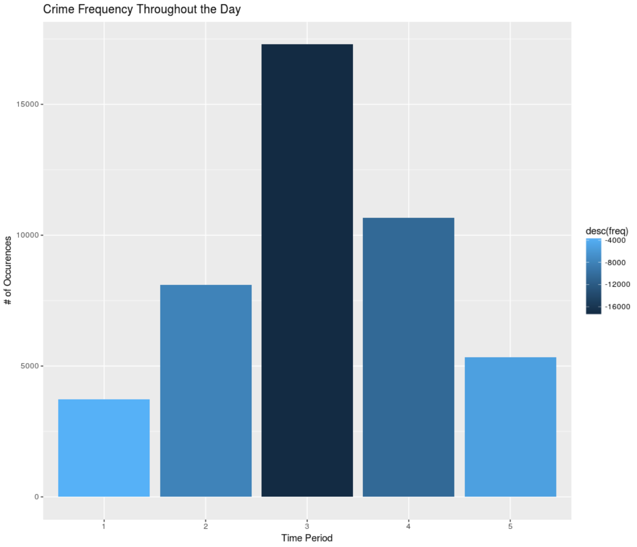

Fig. 3: *Frequency of crimes committed per time period. Each bar is split into the proportion of each crime committed during that hour.*

Finally, instead of looking at the crimes with such a focused lens, we also considered crimes committed throughout the year, grouping the crimes by month. From this perspective, the distribution of crimes during each month is surprisingly uniform, as seen in Figure 5 below. In fact, we found that the difference between the proportions of each crime from month to month to be insignificant. The frequencies of crimes during each of these months were largely uniform as well (See Figure 6) other than the occurrences during December, but this was due to missing data in December over the course of a few years.

### C. Reverse Geocoding Process

We had geographical data in latitude and longitude, but that turned out to be a little too specific for our needs. We wanted to consider geographic trends in larger groups. Thus, we investigated how we could transform these coordinates into something more useful for our task. Zip codes came to mind immediately, and we saw there was a function called `revgeocode()` that calls on the Google API that gave us the information we needed, but we were limited to 2500 queries a day. Evidently, this was not going to work for us.

As a result, we had to create a reverse geocoding mechanism so that we could deal with these queries independently. Adopting a method similar to k-nearest neighbors, we assigned the closest zip code to each of the coordinates. ("Closest" being defined as the smallest euclidean distance from the crime coordinates to the center of a zip code).

```
for (i in 1:nrow(theft_df)) {

  lat = theft_df$lat[i];
  lon = theft_df$long[i];

  if (is.na(lat) || is.na(lon)) {
    next;
  }

  distances = c();
  for (j in 1:nrow(zip)) {
    us_lat = zip$LAT[j];
    us_lon = zip$LNG[j];
    distances[j] = sqrt((us_lat - lat)^2 + (
        us_lon - lon)^2);
  }

  theft_df$zip[i] = zip$ZIP[which.min(
      distances)]
}
```

A small chunk of code used to generate the zip codes is shown above. Once we had zip code, we merged earlier the previously mentioned datasets in 'Data and Variable Description' by zip code.

### D. Analysis

We used the libraries `ggplot2` and `ggmap` to perform a visual analysis of theses variables. After exploring our data visually, we determined which features we wanted to include in a model. Then we decided to use a multinomial logistic regression model and a ranger model, a boosted tree method.

## V. Summary of Findings

### A. Figures

In Figure 4, we can see where in the Los Angeles area `THEFT FROM MOTOR VEHICLE - PETTY` occurs. It appears that these crimes are highly concentrated in the downtown area. Some other geographical visualizations of different crimes can be found in Figures 7 and 8 in the Appendix below.

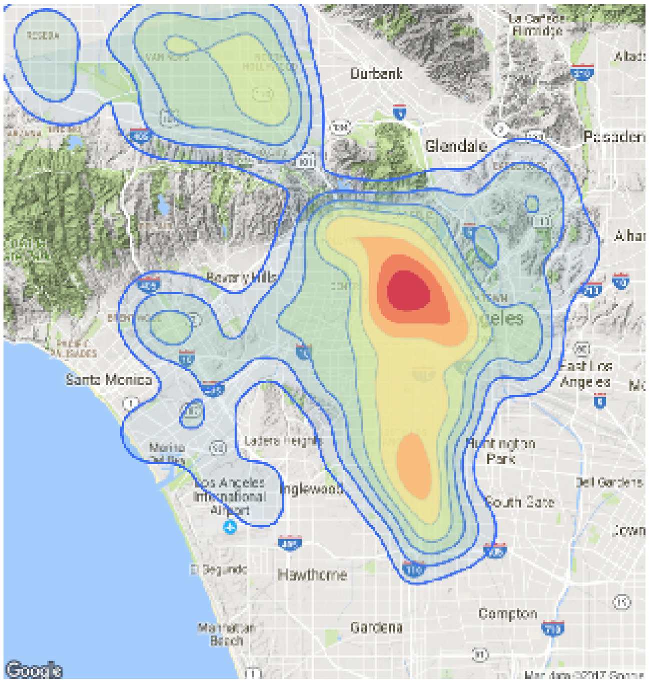

Fig. 4: *Geographical distribution of THEFT FROM MOTOR VEHICLE - PETTY. Red indicates higher density.*

In general, though, we found that the geographical distribution was generally about the same, but there were some variations depending on the crime. In seeing higher concentrations of some crimes over others in certain regions of Los Angeles, we decided perform a quantitative analysis that could shed more light on these visual differences. The motivation behind this was that wealthier suburbs would be more prone to burglaries, while poorer areas may be more susceptible to violent crimes, rather than nonviolent crimes like petty theft. More specifically, we wanted to see variables like wealth, racial/age demographic, etc. affected how/what crimes were committed.

### B. Multinomial Model

With the visual representation of the data in mind, we decided to predict the crime description, `Crm.Cd.Desc` (6 level factor variable), based on the following variables:

- day
- month
- tmp_times
- lat, long
- lat_sq, long_sq
- zip
- Total_Pop
- Median_Age
- Total_Households
- Avg_Household_Size
- Income
- percent_males
- percent_females

In particular, to more precisely model the often nonlinear regional variation in crime incidence, we decided to include squared latitude and longitude.

After splitting our data into a training and testing set, this simple model ended up giving us a very low prediction rate - approximately 18.9%. Note that since there are six groups of crimes we are considering, the prediction accuracy is still slightly better than chance.

```
>> sum(diag(table(predict(model, theft_df),
     theft_df$CrmCd.Desc))) / nrow(theft_df);
>> [1] 0.1897794
```

Investigating further, we saw the exponentiated coefficients of the model were all within 0.9 and 1.2 (this can be seen in an attachment).This made it very difficult to draw any useful conclusions from this model. We suspected the model was not able to interpret the high dimension of the variables, such as zip code which had dozens of levels.

### C. Ranger Model

After suboptimal results from a multinomial model, we used a boosted tree method that is similar to random forests called ranger from library `ranger`. Running the same model predictor and outcome variables as the multinomial model we ended up with better results.

```
>> sum(diag(table(factor(test$CrmCd.Desc),
     response$predictions))) / nrow(test);
>> [1] 0.4015221
```

We were able to double our prediction accuracy to about 40.2%, but this came at the expense of interpretability. Rather than being able to determine the relationship between each factor and the outcome, we were only able to construct an importance table of which predictor variables were important to our model (shown below). Regardless of this fact, the model supports our initial suspicions that time of day and location plays a large factor type of theft being committed.

TABLE I: *Sorted Variable Importance from Ranger Model*

| Variable | Importance |
|---|---|
| tmp_times | 8706.2616 |
| lat_sq | 8380.5732 |
| lat | 833.1343 |
| long | 8194.41 |
| long_sq | 8183.1583 |
| month | 6821.4641 |
| day | 5504.9501 |
| percent_males | 773.0598 |
| percent_females | 733.1143 |
| Avg_HouseHold_size | 712.5166 |
| zip | 661.846 |
| Median_Age | 590.6642 |
| Total_Pop | 575.3216 |
| Total_Households | 565.1277 |
| Income | 388.4376 |

## VI. Conclusions

The initial exploratory analysis with the barplots showed that certain crimes were more common depending on the time of day. The heatmaps suggested that while a wide variety of crime happens generally in one area, there are still some variations in where certain crimes happen most frequently. These conclusions were further supported by the boosted tree model which showed that time and location were important in predicting the type of crime.

Although the results can not be generalized to other areas outside of Los Angeles, they still emphasize a need for more police presence in certain areas at particular times of the day. In areas where a certain type of theft is particularly prevalent or on the rise, having such information can be helpful in better allocating police resources. The models overall were relatively inconclusive, so it is difficult to draw a particularly meaningful cause and effect relationship with just the models. A more refined dataset with more variables may have improved the prediction success of the models.

Understanding time of day's role in theft can be beneficial in the short term because it can help the city of Los Angeles to decide when to have more police officers on patrol. Understanding location's role in theft can be beneficial in the long term because it suggests that development in certain areas may be needed. Although it is out of the scope of this project, community development programs or a stronger emphasis on social work in higher crime rate areas may help to decrease the rate of theft in the long term.

### A. Shortcomings

The public data set does not include information on police activity as well as their allocation of resources, preventing us from rigorously modeling causal relationships. For instance, to examine the intertemporal trends of a particular category of crime over a period of, for instance, two years, we need only the data of each crime incidence and from there we are able to provide aggregate statistics and run regression models. But to examine the effect of police activity on crime, the data on how police activity has changed over time is crucial.

In addition, the absence of police activity data meant we were not able to tackle the interaction of police activity with criminal behavior. In reality, potential criminals respond to police patrol and adjust their activities accordingly. A spike in recorded theft could indicate either an increased incidence of crimes committed or simply an increased police attention to theft. Lacking police activity data, a more comprehensive approach might require borrowing techniques from the studies of economics, game theory or sociology to approximate the effects of this type of interaction.

### B. Recommendations

An explanation of the more unique features of the dataset would have helped to provide more background about the data. For example, some observations in the data had very large discrepancies between the date that the crime was reported and the date that the crime occurred, with some discrepancies ranging up to 500 days. Some more information about the nature of these discrepancies could have also helped with the analysis.

A more refined data collection process would have allowed for more insightful analysis of crime in Los Angeles. Data on officers in Los Angeles such as how many were on patrol on a certain day could have been very beneficial. If more officers are on patrol, it is more likely that crimes will be noticed and reported. However, criminals may also adjust their behavior depending on the number of officers on patrol, which can result in fewer crimes. As a result, the additional information can help with determining if there is a possible relationship between police presence and criminal activity. Additional data such as the age or gender of the criminal could also be beneficial for understanding the relationship between a certain area's demographics and the characteristics of the criminals. For example, if average household income is constant in two areas, but the area that has more schools has fewer juvenile crimes, then there may be evidence that schools are beneficial in reducing juvenile crime.

In terms of improving the model, we could also work on developing a stronger metric. More specifically, we would want to construct one that defines a high degree of "closeness" between similar crimes, and a smaller degree of "closeness" between different crimes. With the final model we built, we were able to determine which variables were "important" in determining the crime description, but this process can be made substantially better with other boosting techniques, such as Adaptive Boost. More training time, in conjunction with a more fastidious method of collecting data could help in constructing a stronger classifier.

## APPENDIX

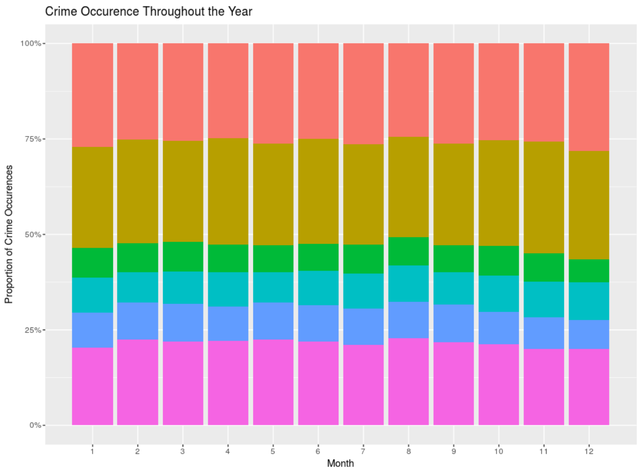

Fig. 5: *Proportion of crimes committed per month of the year. The proportions across each month appear to be relatively uniform, and the differences are in fact insignificant.*

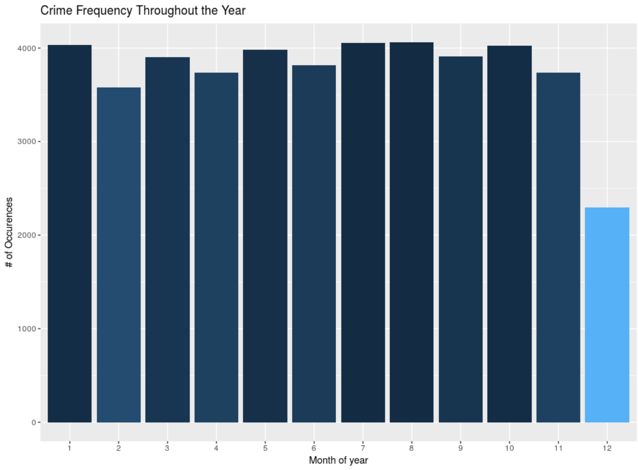

Fig. 6: *The frequency of all crimes committed per month of the year. Even though December appears to have very few reported crimes compared to the other months, but this is because of missing data.*

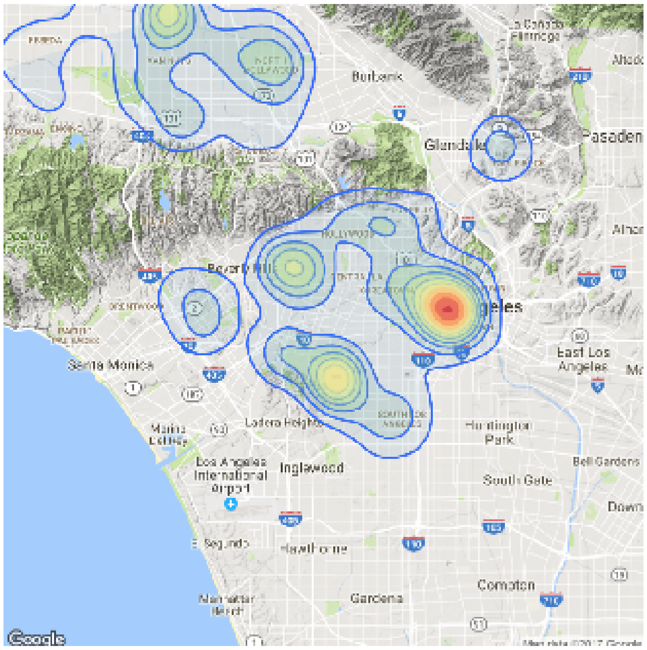

Fig. 7: *Geographical distribution of SHOPLIFTING - PETTY THEFT. Red indicates higher density.*

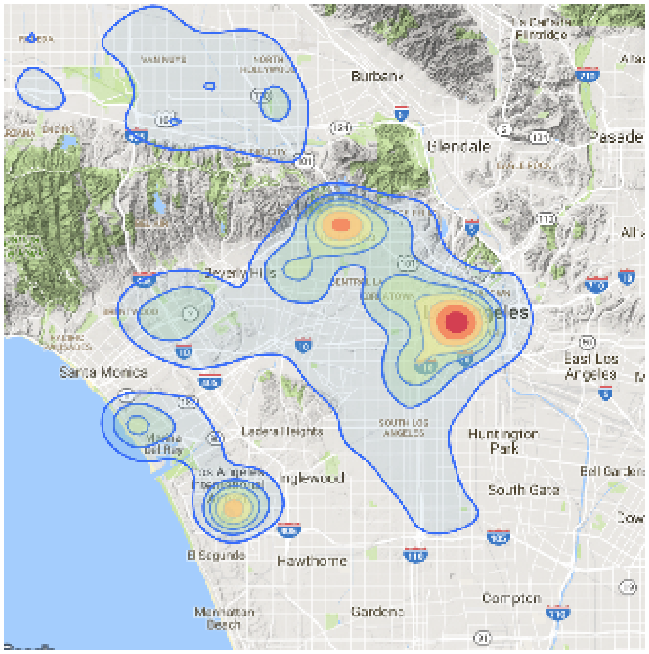

Fig. 8: *Geographical distribution of THEFT - GRAND. Red indicates higher density.*