Impact of Weather and Unemployment on Crime in LA

Eoin Doherty
University of Colorado: Boulder
Boulder, CO, USA
eoin.doherty@colorado.edu

James Maxwell
University of Colorado: Boulder
Boulder, CO, USA
jama4534y@colorado.edu

Ryan Murphy
University of Colorado: Boulder
Boulder, CO, USA
rymu8236@colorado.edu

Ryan Shuman
University of Colorado: Boulder
Boulder, CO, USA
ryan.shuman@colorado.edu

ABSTRACT

We propose to investigate to what extent Weather and Unemployment impact crime in Los Angeles.

ACM Reference Format:
Eoin Doherty, James Maxwell, Ryan Murphy, and Ryan Shuman. 2018. Impact of Weather and Unemployment on Crime in LA . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

We are interested in answering a variety of questions including: What types of crimes are committed the most? When are most crimes committed? What areas of LA city have the most crime? Has the crime rate in those areas improved since 2010? Is there a correlation between weather and crime in LA? Unemployment rate and crime? If so, how much of an effect do unemployment and and weather have on crime rates and types of crime? We believe that this type of analysis could allow the LA police force and potentially other police forces to better allocate resources. Identifying when and where crimes are more likely could allow officers to be in position to respond to crimes more rapidly. For instance if we find that on days that are hotter than average a greater proportion of crimes occur in district 2 relative to on days when it is cooler it might make sense to station a larger number officers in district 2.

2 LITERATURE SURVEY

In a kaggle competition the LA crime dataset was used. The competitors noticed some interesting things such as assaults and vehicle thefts happen mostly in the evening while petty thefts seem to happen around noon and burglaries happen when people aren't at their homes(in the morning to afternoon). They also noticed a upward trend in robberies and vandalism. They also noticed a correlation between similar types of crimes. For example burglary from vehicle and vehicle thefts tended to occur frequently close together prompting the observer to think there may be "external causes that drive similar crimes (weather, general mood, ... don't know)". It would be interesting to see if weather or economics are these "external causes".

Previous studies performed in the UK "...yielded strong evidence that temperature has a positive effect on most types of property and violent crime. The effect was independent of seasonal variation. No relationship between crime and rainfall or hours of sunshine emerged in the study." We are interested in seeing if similar results are found in Los Angeles. LA experiences relatively mild temperature variation which may allow us to isolate other relationships.

Prior research on the subject of unemployment and crime rates have yielded mixed results so it will be interesting to see what evidence we find.

3 PROPOSED WORK

Our data are pulled from multiple sources. The data from these sources have null values and extraneous information, so we will need to clean it thoroughly before we can start analyzing. We will also need to consolidate the data from our multiple data sets into one dataset to make analysis easier.

First, we will need to drop some values from each individual dataset. The unemployment data contains monthly statistics from January 1990 to December of 2017, but since the crime data we have is from 2010, we will drop all data before then. Since the unemployment rate data ends in December of 2017, we will have to drop the data from 2018 in the crime data set. The weather data also contains some data from 2018 that we will have to drop.

We will also need to drop some extraneous data from each dataset. The unemployment rate dataset is only two columns, with no empty entries, so we will not need to clean it further. The crime dataset could use some cleaning however. It has some columns that are useful to a human reading the data, but take extra memory when analyzing the data. We will have to drop the "Crime Code Description" column, the "Premise Description" column, the "Weapon Description" column, and the "Status Description" column from the data set since they correspond to other columns that contain codes that are easier to process. To keep track of what those codes mean, we will create a short reference csv that will not be included in data analysis but can be used to make sense of the results of our analysis. The weather data also contains a lot of extra information that we will not need. We will not need the "STATION", "STATION_NAME", "ELEVATION", "LATITUDE", and "LONGITUDE" columns since

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

all of the data was taken from the same weather station in downtown LA. We are also not getting too specific with the weather, so we will not need "REPORTTPYE", "SKYCONDITIONS", and the columns referring to pressure or altimeter setting. The weather dataset was generated by a computer from a larger dataset, so we will have to drop a lot of columns that are empty or mostly empty like the columns for monthly weather. Lastly, we will have to drop all of the temperature columns in Celsius since we will only use the temperature in Fahrenheit.

Some of the entries in our datasets have been left blank or have inconsistent values. While we have dropped some mostly empty columns before, useful information can still be gained from some of the entries that have been filled out. To make processing the data easier, we will need to fill empty entries with a default value. In the crime data, we will fill empty entries in "Victim Age" with -1 since that is an integer that will be easy to filter. For "Victim Sex" and "Victim Descent", we will add the character '-' as a placeholder since those columns have one character per row. Finally, for "Crime Code 2" to "Crime Code 4", we will replace empty values with -1 since that does not correspond to a real crime code. The crime data's location field is a tuple of latitude and longitude, so we will split it into two columns: one for latitude and another for longitude. The weather data contains hourly weather data followed by a summary of the daily weather data at 11:59 PM every day. We will split the weather data into two datasets. One for daily weather and one for hourly weather. We will drop the hourly fields from the daily weather data set and drop the daily fields from the hourly weather data set. We will also have to fill in empty entries in hourly wind speed, and daily snow depth with 0.

After our data has been cleaned, we will join the datasets together to form one dataset that is easy to process. Each row of this combined data set will correspond to a crime from the LA crimes dataset. We will add columns for the hourly and daily weather for when the crime was committed to the end of the row and add a column for the unemployment rate for that month.

Once our data has been consolidated, we can analyze it. We will start by analyzing the crime data overall. We will find the frequency of crime as a function of time and the frequency of crime as a function of location. We will also find the most frequent types of crimes and where they occur. When calculating the areas where crimes occur, however, we should also adjust for the number of people living in that area since crime rate is likely higher in more densely populated areas. One crime can have multiple crime codes, so we can generate frequent patterns with an apriori algorithm to see which crimes occur together most often. We can also find statistics about the victims such as their average age, their most frequent gender, their most frequent descent, or even the average amount of time it takes them to report a crime.

Once we have done this simpler analysis, we can start to mine for more interesting patterns and correlations. We can use data on the monthly frequency of crimes to see if there is any correlation with unemployment rate since we assume crime would increase when more people are unemployed. We can also see if there is a correlation between good weather and crime rate by comparing crime frequency to heat index, precipitation, and visibility for each day. We assume there will be more crimes committed on days with good weather. We can also use the sunrise and sunset times from the weather dataset to see if crime rate increases after dark or when nights are longer. We can group our data by crime type to look for correlations between crime type and weather or crime type and unemployment rate since people may steal more when it is cold outside or when the unemployment rate is high. Our project is not revolutionary. There is research on how unemployment rate or weather affect crime rate, but we would like to see if we can replicate these results on a smaller, one city scale. Furthermore, our analysis takes unemployment rate, weather, location, and many other factors into account instead of focusing on one variable's effect on crime rate.

4 DATA SET

We will use data from three different categories.

4.1 Crime Data

LA Crime Data from 2010 to Present
https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq
The LA Crime Data from 2010 to Present is composed of of 26 columns which detail the date and time a crime occured the date it was reported the type of crime (as a text field) Victim details including sex, age and descent, a modus operandi field that details specific actions the perpetrator was suspected of committing (numerically coded the corresponding lookup table is MO Lookup table), the police district that the crime was committed in (numbered 1 through 21), the address the crime occurred at (text), a premise description that details the specific surroundings of the crime (categorical eg sidewalk, Park/Playground, Market, apartment etc). Weapon details including a numeric weapon code with associated lookup table, weapon description(categorical text eg Knife with blade over 6 inches in length, strong arm(hands, fist, feet or bodily force etc) 4 numeric crime codes that detail the specific crime in decreasing significance (eg code 1 is most significant, codes need to be cross referenced and a geographic code that is geotagged at a 100 block accuracy level.

4.2 Weather Data

NOAA Weather data for LA City 2010 to present
https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

The NOAA Weather data for LA City 2010 to present, details hourly weather measurements for downtown LA including Station id, Station name, latitude and longitude, Date, Time Temperature,Precipitation, Air Pressure, Sunrise and Sunset. The data is numeric but has many holes, precipitation for instance is sometimes simply blank and other times a 0 entry has been entered these inconsistencies will need to be corrected before we can proceed.

4.3 Unemployment Data

LA Unemployment Rate from 1990 to December 2018
https://fred.stlouisfed.org/series/CALOSA7URN

The LA Unemployment Rate from the Federal Reserve lists monthly unemployment rate for LA from 1989 through the present. Unemployment is reported as a percentage eg (5.7,7.9 etc).

5 EVALUATION METHODS

We will use a number of standard metrics in our analysis including:

5.1 Heat Index

Heat index provides an elegant means to aggregate weather data and provide a closer approximation to human experience:

$$HI = c_1 + c_2T + c_3R + c_4TR + c_5T^2 + c_6R^2 + c_7T^2R + c_8TR^2 + c_9T^2R^2$$

where HI = heat index (in degrees Fahrenheit)
T = ambient dry-bulb temperature (in degrees Fahrenheit)
R = relative humidity (percentage value between 0 and 100)
$c_1 = -42.379$, $c_2 = 2.04901523$, $c_3 = 10.14333127$, $c_4 = -0.22475541$, $c_5 = -6.83783 * 10^{-3}$, $c_6 = -5.481717 * 10^{-2}$, $c_7 = 1.22874 * 10^{-3}$, $c_8 = 8.5282 * 10^{-4}$, $c_9 = -1.99 * 10^{-6}$

5.2 Crime Rates

Police District Crime Rate/Neighborhood Crime Rate: crimes per 100,000 population per year (or other unit time but per year is customary) Can also be specified by type of crime (eg murder rate, violent crime rate, burglary rate etc.)

5.3 Unemployment

We failed to find a standardized metric for relating unemployment and crime however it is easy to conceive of crimes per percent unemployed per year.

6 TOOLS

We will need to use a few tools to clean and analyze our data. Since the data we are using comes from multiple sources and is therefore fairly dirty, we will use MySQL to clean and aggregate it into the datasets described previously. After the data has been cleaned and aggregated, we will use python and pandas to mine it for interesting patterns. We will use jupyter notebooks for our python data analysis since they are efficient and easy to use. Since we are using python, we will also use matplotlib to generate plots and visualizations of our data that can be used to analyze and present our findings.

7 MILESTONES

Activity	Date	Completed
Begin writing Project Proposal	03/01/18	Y
Submit Project Proposal	03/06/18	Y
Begin Data Cleaning	03/08/18	Y
Complete Data Cleaning	03/15/18	Y
Begin Data Consolidation	03/16/18	Y
Complete Data Consolidation	03/21/18	Y
Begin Data Analysis	03/22/18	Y
Complete Frequency Analysis	03/8/18	1/2
Write Progress Report	04/8/18	Y
Progress Report Due	04/10/18	Y
Complete Data Analysis	04/12/18	Y
Evaluate Results	04/13/18	Y
Begin creating visualizations	04/14/18	Y
Finish visualizations	04/15/18	Y
Begin Presentation Development	04/18/18	Y
Finish Presentation	04/22/18	Y
Begin Final Report	04/23/18	Y
Finish Final Report	04/27/18	Y
Final Project Due	05/01/18	Y

7.1 Pre-Processing

Cleaning: First, we dropped some values from each individual dataset. We dropped unemployment data prior to 2010 since our crime data starts in 2010. We also dropped data from 2018 in the crime data set and weather data set since our unemployment rate data ends in December of 2017.

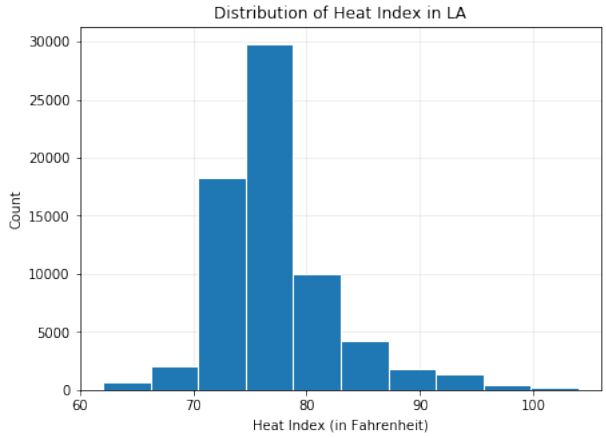
From the crime dataset we dropped the "Crime Code Description" column, the "Premise Description" column, the "Weapon Description" column, and the "Status Description" column from the data set since they correspond to other columns that contain codes that are easier to process. To keep track of what those codes mean, we created a short reference csv that will not be included in data analysis but can be used to make sense of the results of our analysis. In addition we dropped many unnecessary columns from the weather dataset. We dropped the "STATION", "STATION_NAME", "ELEVATION", "LATITUDE", and "LONGITUDE" columns since all of the data was taken from the same weather station in downtown LA. We are also not getting too specific with the weather, so we got rid of "REPORTTPYE", "SKYCONDITIONS", and the columns referring to pressure or altimeter setting.

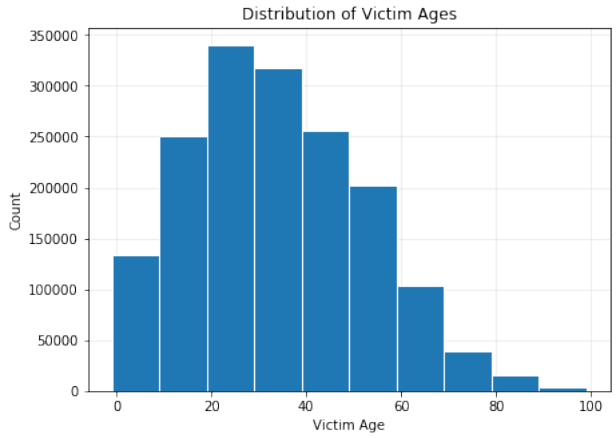
To allow for proper analysis we filled in missing values as described in section 3.

8 RESULTS

Here is some basic analysis we did that yeilded interesting results:
Averages:
victim age: 32.97893772232137

Average unemployment rate: 8.649744172380776
Average Heat Index: 77.65726601980991

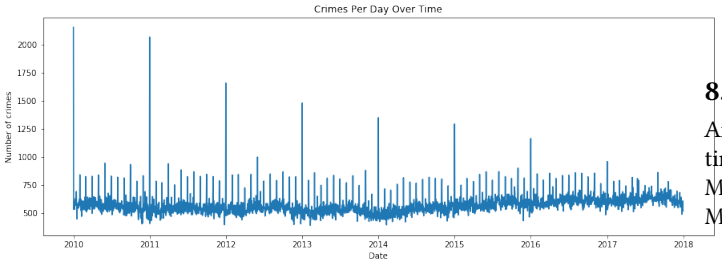




8.1 Frequencies

The most common place crimes in LA occur is in order: Street, Single Family Dwelling, Multi-Unit Dwelling, Parking Lot etc.

The most common weapon was by far STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE).

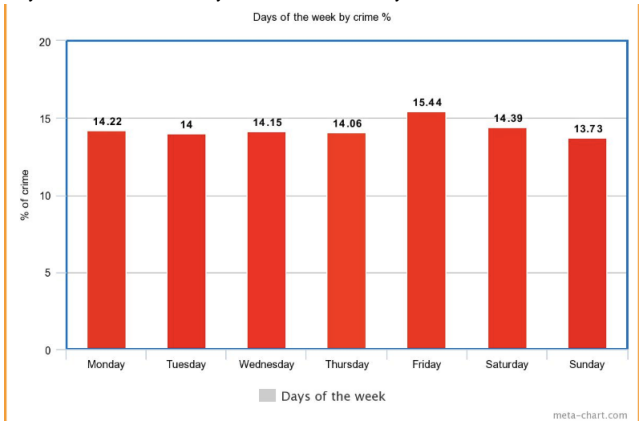


Looking at the above graphic there are major spikes at the first of the year and more minor spikes at the 1st of each month.
There were some interesting observations made on dates. There doesn't seem to be a significantly more crime when split up by the months of the year. However something interesting happens when looking at what day of the month these crimes occur at. According to the data on the first day of the month criminals come out in full force and wreak havoc on the community. This we believe is just noise possibly due to crimes without a know occurrence date being logged to the first of the month. This may need to be corrected in future analysis. There are also less crimes on the 31st of the month but this is due to many months not having 31 days.

Another thing that was interesting when looking at dates is seeing the distribution of crimes based on days of the week. For days of the week the crime was reported the distribution is: Day Monday 15.500541 Tuesday 15.242425 Wednesday 15.013330 Thursday 14.779178 Friday 14.387519 Saturday 12.680597 Sunday 12.396411

From the start of the week on monday to Sunday there is a steady decrease of crimes reported with it being the lowest on the weekend.

The date the crime occured breaks down as follows: Friday 15.440332 Saturday 14.391794 Monday 14.228808 Wednesday 14.151199 Thursday 14.061668 Tuesday 13.999533 Sunday 13.726666



It makes sense that the sleepiest day of the week(Sunday) is also the safest.

8.2 Time taken to report crimes

Another thing we found that was interesting was the amount of time it took to report crimes.
Mean: 16.192 Days
Median: .083333 Days (massive skew)

Shortest time to be reported by crime:

Rank	Crime
1	BATTERY - SIMPLE ASSAULT
2	BURGLARY FROM VEHICLE
3	VEHICLE - STOLEN
4	BURGLARY
5	THEFT PLAIN - PETTY ($950 & UNDER)

This makes sense since these are common crimes that the victim knows about immediately.

Longest time to be reported by crime:

Rank	Crime
1	THEFT OF IDENTITY
2	THEFT-GRAND
3	THEFT PLAIN - PETTY
4	DOCUMENT FORGERY / STOLEN FELONY
5	BURGLARY

Intuitively this makes sense as well. It can sometimes take years to figure out your identity has been stolen. Same idea with document forgery.

8.3 Where do the crimes occur?

The data set splits the crime locations into 21 possible areas across LA. These locations are how the police are split up.
Top five most violent areas are:

Area	% of Total Crimes
77th Street	6.967615
Southwest	6.452347
N Hollywood	5.439272
Pacific	5.286643
Southeast	5.265449

With most of these crimes occurring on the streets or in houses:

Location	% of Total Crimes
STREET	22.273922
SINGLE FAMILY DWELLING	20.633349
MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	12.869291
PARKING LOT	7.117053
SIDEWALK	4.992341

8.4 What crimes are commited most?

Some crimes have multiple crime codes.We used apriori pruning with a minimum support of 0.0003 to find frequent itemsets of crime codes.
Most frequent 1-itemsets:

Crime
BATTERY - SIMPLE ASSAULT
VEHICLE - STOLEN
BURGLARY FROM VEHICLE
BURGLARY
THEFT PLAIN - PETTY ($950 & UN

These crime codes can apply to a variety of crimes and are rather ambiguous.

Most frequent 2-itemsets:

Crime 1	Crime 2
CRM AGNST CHLD (13 OR UNDER)	BATTERY WITH SEXUAL CONTACT
BRANDISH WEAPON	CRIMINAL THREATS - NO WEAPON D
THEFT PLAIN - PETTY ($950 & UN	BIKE - STOLEN
THEFT OF IDENTITY	DOCUMENT FORGERY / STOLEN FELONY

These get a little bit more specific.

8.5 Miscellaneous Statistics

Victim Stats:

Sex	% of Victims
M	46.603255
F	42.544633
Not Reported	10.852112

Descent	% of Victims
Hispanic	34.641071
White	24.644707
Black	16.044107
X (Unknown)	11.955079
Other	9.658110

Weapon Stats:

Weapon	% of Total Crimes
STRONG-ARM (BODILY FORCE)	20.148848
VERBAL THREAT	2.750222
UNKNOWN/OTHER WEAPON	2.584226
HAND GUN	1.595957
SEMI-AUTOMATIC PISTOL	0.630869

8.6 Correlations

Slight negative correlation of -0.141 between the unemployment rate and number of crimes committed for a given day.

Slight negative correlation of -0.013 between heat index and the total number of crimes in a day. Discretizing this dimension did not help.

Very slight positive correlation of 0.002 between crime and humidity. This probably isn't very useful.

Slight negative correlation of -0.035 between precipitation and crime. Precipitation had a significant effect on crime rate when days were grouped as rainy or not rainy.

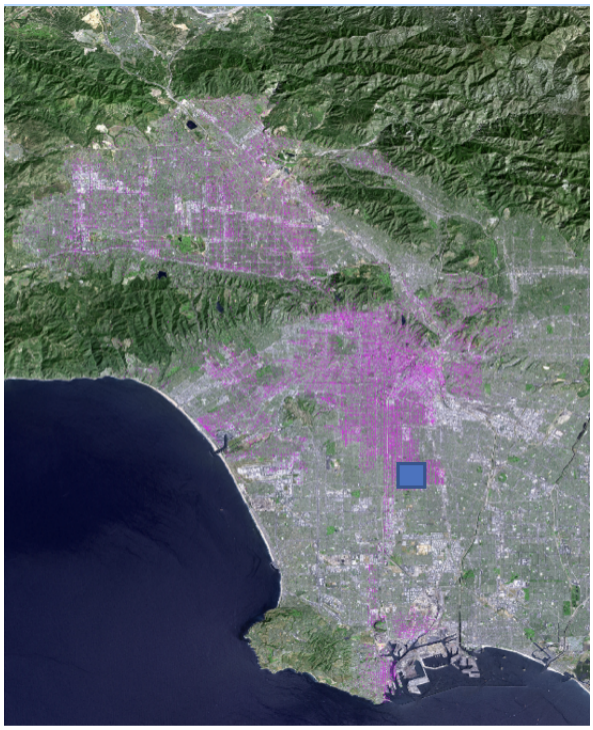
Rain	Average Number of Crimes (Daily)
No	540.1561829878484
Yes	349.3621495327103

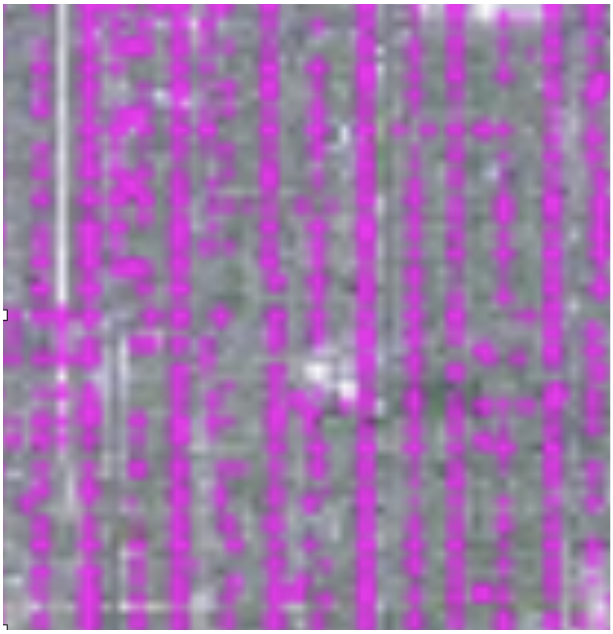
There is a fairly significant negative correlation of -0.529 between crime and unemployment rate. This is counter-intuitive to what we would believe so we think there may be another variable affecting both crime and unemployment rate.

8.7 Clustering

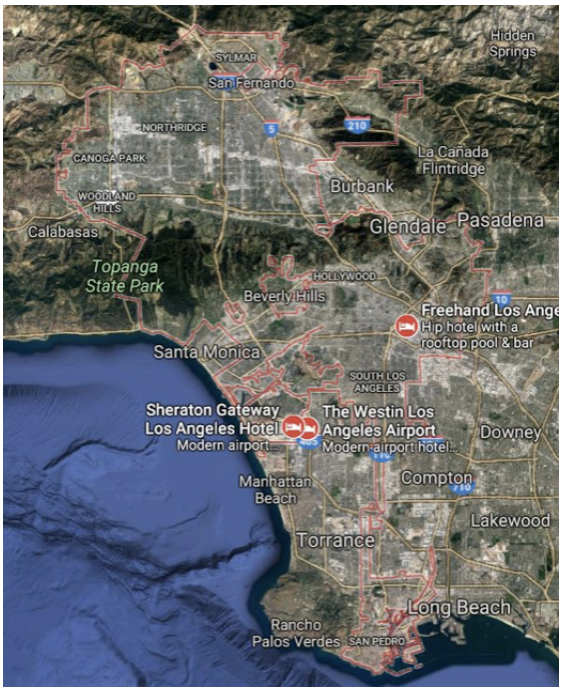
The purpose of clustering in our project was to detect possible correlations between the latitude, longitude locations of each crime and other categories they fall under. In particular, to discover a possible connection between rain, unemployment rates, and the locations of crimes. To accomplish this, we used DBSCAN from python's SKlearn module.

Our first concern was that the accuracy, or precision of the latitude and longitude would be too inadequate to be reliable. As for precision, the location data was almost entirely to the 4 decimal place, since every degree latitude or longitude is 60 nautical miles therefore most location data was accurate to .006 nautical mile or roughly 36 feet, since LA is 503 square miles, we deemed 36 feet to be precise enough to proceed. The next step was to make sure that the data was accurate. To solve this, we mapped every datapoint over a large map of LA. First is the full picture and second is the zoomed in blue section.





As a reference here is a map of LA.

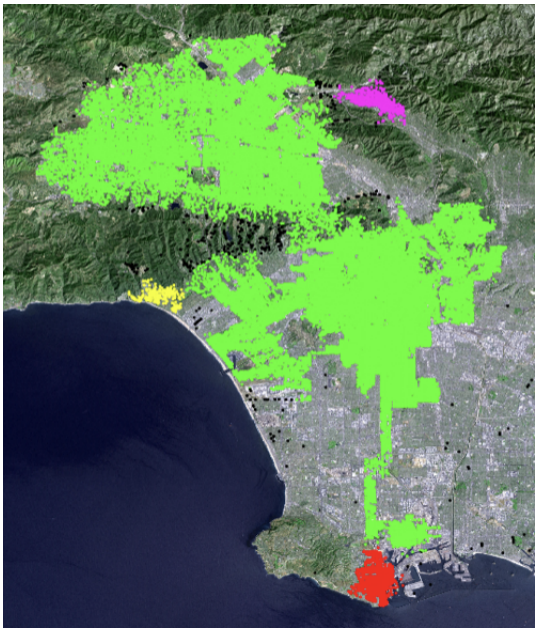


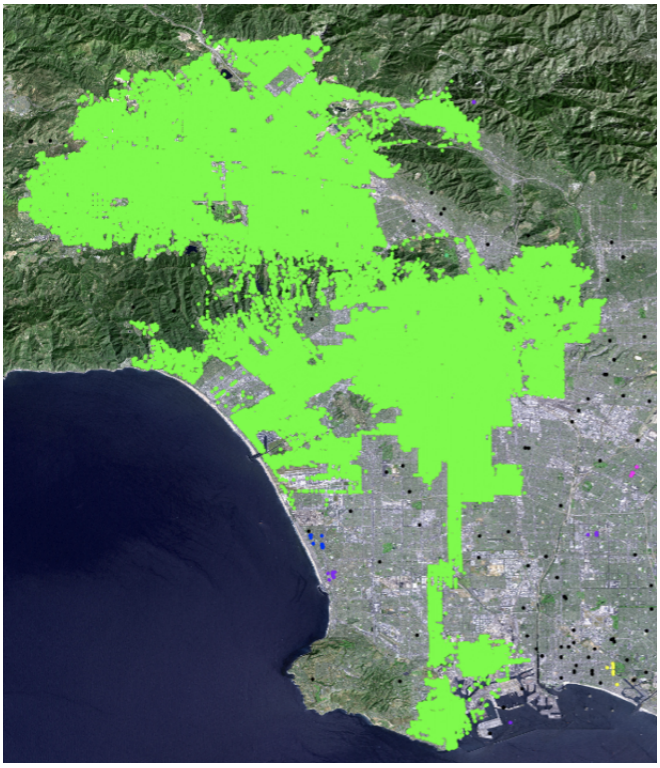
Given that the data we can see line up with streets very accurately, as well the outline of all the data matches the outline of the LA Police Office's jurisdiction very well we were confident that the data was accurate and that we could use it for clustering.

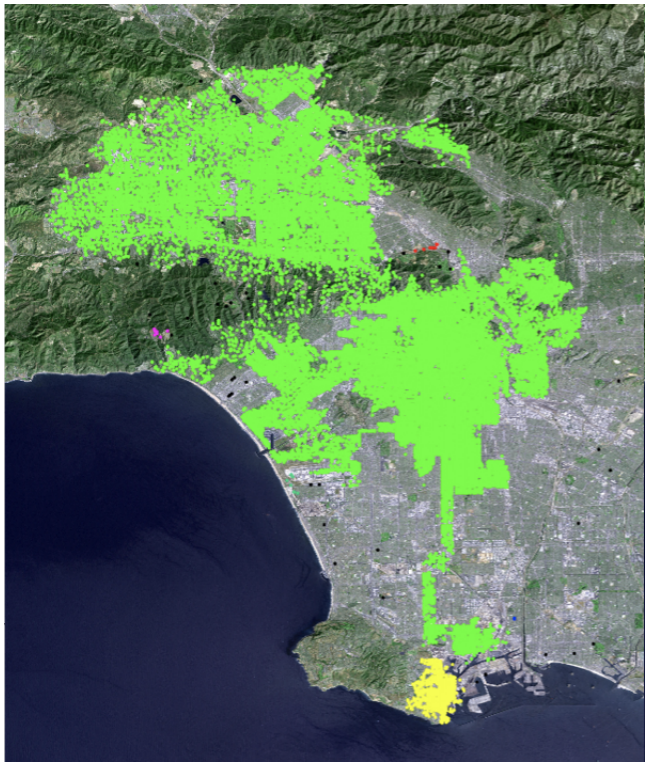
The way we decided to perform our clustering was to isolate the data with particular attributes we wished to test. We would then cluster based on that newly isolated data to see if we got any significant clustering. For our minimum distance for clustering datapoints we tested many possible numbers and used different numbers for different attributes, but found the best results when we used .01 degrees Latitude, longitude, which is about .7 miles which makes sense since each city block in LA is roughly 1 square mile. The number had to be large enough to give us significant clusters, without discounting possible ones. The max number of samples we used varied more than the minimum distance for each attribute. We would test many for each set of attributes and use the one that gave us the best result. It was commonly 1/500 of the total dataset length. Something we also had to consider was the curvature of the earth, we decided not to use a haversine calculation for our scans, and instead rely on Euclidean. The thought was that our data represents an area with a length slightly less than 45 miles max along the North South Axis. This represents about 1350 feet of distance that would be skewed due to the earths curvature. We felt this was negligible given the density of our dataset.

We ran into some initial problems with the clustering. The most time consuming to deal with, was the issue of the data itself. Some of the data was 0,0 for the latitude and longitude, some of the data was blank, and some were just nonsense symbols. Luckily, this represented a very small amount, so we decided to omit this data from our clustering. In total we had 6149 omissions. The second issue we had was that DBSCANS, indeed, clustering in general, is not usually performed on such a large dataset. We performed the operations on a computer with plenty of processing power and 16 gb or RAM but were only able to scan over a dataset of about 400,000 points before running out of memory. We tried normalizing the data, and preprocessing but neither were able to make any significant dent in the memory usage, most likely due to the nature of our dataset having one distinct cluster with points that are quite evenly spaced, with only a tiny bit of noise. We got around this by only choosing attributes that gave us a dataset less than 400,000. That way we could still run over the entire dataset, but still be able to compute the clustering. Another unforeseen issue we had with our data was that because we had data that was very dense, it was very ineffective at finding subtle correlations, only those correlations that were very pronounced stood out.

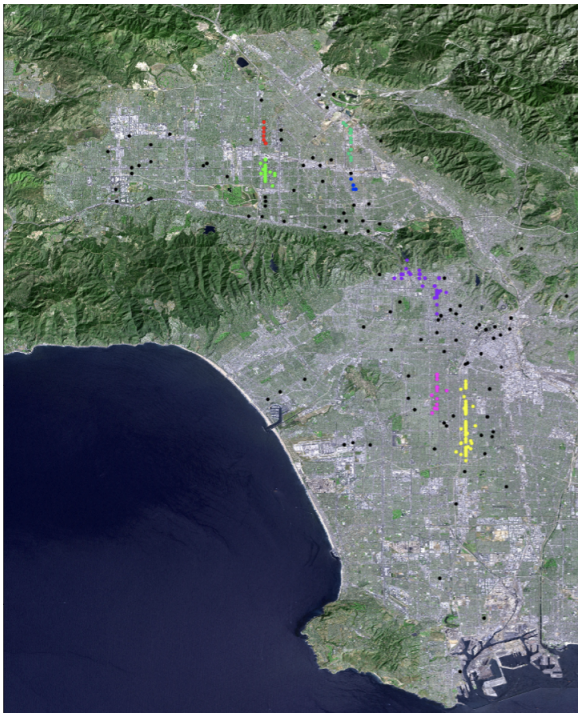
Our primary objective was to see if rain, or unemployment rates effected the location of where the crimes were committed. We clustered low unemployment, high unemployment, crimes committed on days it rained, crimes committed on days it didn't rain; as well as, high unemployment with rain, low unemployment with rain, high unemployment with no rain, low unemployment with no rain. None gave any significant clusters. The first one is all crime on days without rain, second is all crime with rain, and third is employment rates. The clusters that appear on the first and second are not statistically significant and a result of plotting less points in order to effectively run the dbscan.



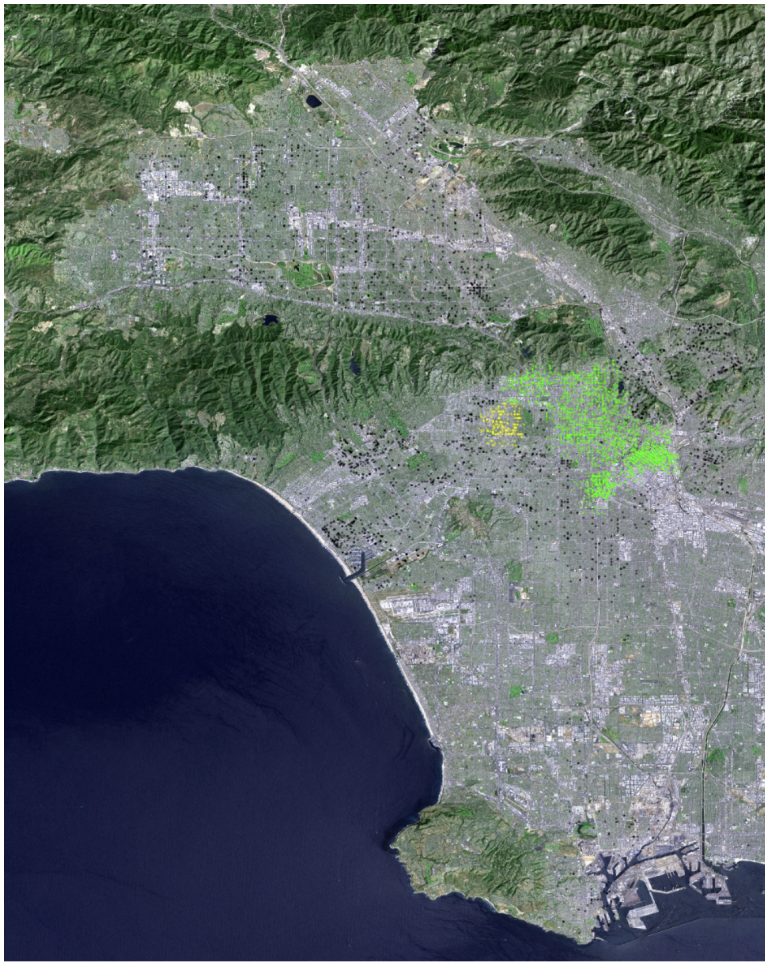




Since we could not find any significant clusters between unemployment rates, and rain, we decided to search for other significant clusters. After an exhaustive search, we found two datasets that gave significant clusters. The first was clustering based on the crime code associated with pimps.



The clustering clearly shows a connection between pimps and location what's interesting about this data is that all but 1 of the clusters line up north to south along a single street. Further exploration shows that the lines occur at noticeable intervals of time. First the yellow line partially appears then much later the magenta, then the red, green, and blue in the valley, and finally the purple. Upon further investigation, it seems that the police departments every once in a while, round up pimps all at once which means the lines are probably less due to the probability that pimps only hang out on north-south lines but rather that the police will do a sweep along 1 street before the pimps disperse. However, this is just a hypothesis. The second interesting information we gained from clustering was looking at stolen bicycles.



There is a significant cluster of stolen bicycles around the downtown area, extending into eastern Hollywood. There is another separated cluster around west Hollywood. A bit of this clustering is easily explainable, if you look at the green cluster there are two thicker areas to the east and south, we tried to get these to separate into their own clusters, unfortunately there were too many points in between to fully separate them, but they make sense because the one to the east is downtown, and the one to the south, perfectly encircles the USC campus. We were unable to figure out what the significance of the yellow cluster was, however some research did suggest that bikes are commonly stolen from people's garages in that area. Something else of interest not necessarily shown by the clustering but by the mapped data itself, was the lack of bike lefts in the south LA area. Typically, south LA has a much higher crime rate than the other parts of LA, but not in bike thefts. Our hypothesis is that bike thefts are not as reported in south LA as they are in the more affluent areas, such as the area within the yellow cluster.

8.8 Application of knowledge

There are several things that we can take from what we learned. Police and government officials can more intelligently plan police location staffing and routes to more accurately account for crimes that are occurring. For example with prostitution going on on certain streets it would be appropriate for an officer to be checking those specific areas. On top of that the 77Th st area seems to be very crime ridden so city officials may want to allocate more money towards schools there. It would seem less police are needed on rainy days. Another thing is that it takes awhile for victims to report some crimes like identity theft. The city of LA could see something like this and decide to educate people on how to spot identity theft.

REFERENCES