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, 3 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 âĂIJexternal causes that drive similar crimes (weather, general mood, ... don't know)âĂİ. It would be interesting to see if weather or economics are these âĂIJexternal causesâĂİ.

Previous studies performed in the UK âĂIJ...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.âĂİ 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",

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

"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 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 variableâĂŹs 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	
Complete Data Cleaning	03/15/18	
Begin Data Consolidation	03/16/18	
Complete Data Consolidation	03/21/18	
Begin Data Analysis	03/22/18	
Complete Frequency Analysis	03/30/18	
Write Progress Report	04/01/18	
Progress Report Due	04/03/18	
Complete Data Analysis	04/10/18	
Evaluate Results	04/11/18	
Begin creating visualizations	04/12/18	
Finish visualizations	04/14/18	
Begin Presentation Development	04/18/18	
Finish Presentation	04/22/18	
Begin Final Report	04/23/18	
Finish Final Report	04/27/18	
Final Project Due	05/01/18	

8 SUMMARY OF PEER REVIEW SESSION

The most significant feedback that we received was that we needed to provide greater detail about what procedure we would use for cleaning, organizing and analyzing our data. We also were encouraged to be more specific about what we hoped to find and how we would use our findings. We have addressed both of these questions in our proposal. Our proposed work section explains in great detail how we plan to pursue our analysis. In our problem statement you will find an explanation of the kinds of questions we plan to ask and potential uses for the answers.

REFERENCES