Crime Time, Together with Weather

[Proposal]

Jackson Markowski
University of Colorado
jama1181@colorado.edu

Jayce Meyer
University of Colorado
jame2714@colorado.edu

Miles Shayler
University of Colorado
mish8391@colorado.edu

Conor Walsh
University of Colorado
cowa6926@colorado.edu

1. PROBLEM STATEMENT

The motivation behind our project is simple: to discover the correlation, if any, between crime, weather, and the time period of the occurrence. We want to take a deeper look at the data for these things in order to understand any patterns that may exist within them. Could a warm, sunny day make it more likely for crime to be committed? Is crime more frequent during the snowy winter months when the days are shorter? Do crimes occur more frequently during specific times of the year, such as holidays? These are all questions we intend to answer using the data we have found.

In addition to answering these, we also want to form analytical theories as well. For instance, if we find a tie between summer and high crime, we would theorize what reasons there might be to commit a crime when weather is typically nicer. Our goal is to uncover the strongest and most pronounced correlation or pattern in our data.

2. LITERATURE SURVEY

Crime is a major aspect of any society. This leads to it be heavily analyzed by cities and law enforcement agencies in order to get the best idea of when and where crime occurs. Most cities have a fairly good understanding of where high crime areas are, which crimes are occurring, and when everything is happening. Like most cities, Denver, has statistical information such as a heat map of crimes and their locations[3]. All this information is then utilized in order to better predict and understand crime.

In terms of weather and its role on crime, many people have heard of some type of environmental condition potentially having some effect on crime. Intuitively, most people might assume less crime would occur when its the middle of winter and freezing outside. An article in the New York Times, which looked at a couple dozen studies regarding conflict and weather, found that there is a correlation between conflict and higher temperatures[1]. Very similar
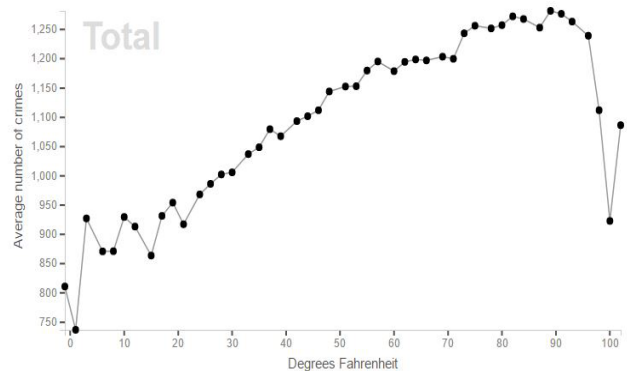
Figure 1: Average number of crimes vs temperature (crime.static-eric.com)[6]

statistics were also found in a study for crime in Chicago from 2001 which graphed temperature data against criminal activity[6]. However, what is particularly interesting about this study is that while things like total crime increase with average temperature (**Figure 1**), when looking at individual crimes, such as prostitution, it's not guaranteed to increase with temperature (**Figure 2**).

It has also been frequently hypothesized that things like the moon cycle could have an effect on crime. While it still is debated on if the Moon actually effects society, there have been studies that have found higher levels of crime occurring on full moons as opposed to other lunar phases[5].

3. PROPOSED WORK

In order for us to begin data collection we must first pre-process our two data sets. Combined, our data sets have an immense amount of rows. If we wish to find correlations between crime, weather, time, and locations we will need to first clean our data. The first step to cleaning our data will be to remove the null/missing values, especially in our crime data set. Redundant data is also very prevalent within our data sets, specifically dates and locations. To account for this redundancy, we will remove data that is repeatedly occurring that doesn't help us and only slows down the process. The last issue we need to tackle in order to clean our data is the removal of noisy data. Throughout our records we have a few duplicate records, and the occasional row with
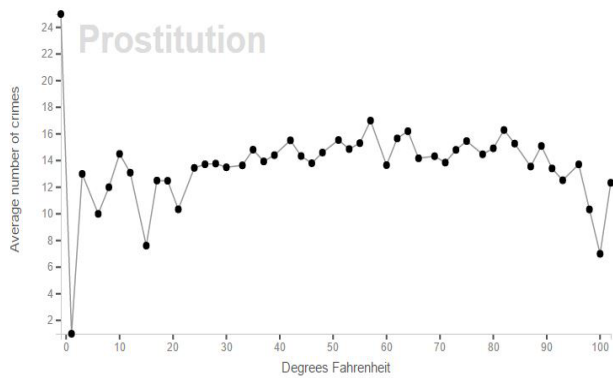
Figure 2: Prostitution crimes vs temperature (crime.static-eric.com)[6]

incomplete data. We will also remove attribute types that aren't helpful to us, allowing us to narrow down our data further.

The next step we propose to do in order to begin finding correlations requires us to transform our data. To make our jobs easier, we will first rearrange our data sets into chronological order. Then we will begin to group the data by specific crime type (for our crime data set) and the location (for both sets). By arranging our crime data set in chronological order, crime type, and location we will be able to easily compare it with our weather data set. We may even add a new attribute type that combines both the crime type and location (only applicable for the crime data set). Doing this would reduce the number of attributes we would have to work with.

To begin finding correlations between our crime and weather data sets, we must integrate the two sets. By rearranging our data chronologically, we can easily match the weather on the day of the crime we are looking at. One of the issues we must handle when integrating our data is that we must match the location of the crime with the location of the weather. If the location of the weather is to general, then we cannot necessarily find a correlation between the weather and the crime.

4. DATA SETS

Our analysis will primarily consist of utilizing two data sets. One of which is criminal activity and the other is weather. We intend to focus exclusively on Denver, Colorado. This ultimately means our data and analysis will only be for the Denver Metro area.

4.1 Crime Data

The crime data comes from a data set[2] of crimes reported in the City and County of Denver. It is provided by the Denver Police Department for public use. To insure its accuracy the data is constantly being updated. This includes additions, deletion, and any possible modifications to entries. Due to legal reasons, the data does not include crimes involving juveniles and does not include location data for sexual assaults.

For the purpose of our analysis we intend to use the data from the start of 2012 to the start of February 2017. Due

to the data being dynamic there is a possibility that some of the data we use could change. However, since most of the entries tend to remain unchanged and the data set includes roughly 500,000 incidents, our analysis should remain accurate regarding the potentially changing data.

As seen in **Table 1**, the attributes for the crime data includes the specific type of crime(homicide, trespassing, etc.), dates regarding the incident, as well as general location information.

4.2 Weather Data

The weather/environmental data set[4] comes from National Centers For Environmental Information(NOAA). The data was taken from NOAA's "Daily Summaries" data set corresponding to the same time range as the crime data. Each day includes entries from multiple weather stations around the Denver area.

As seen in **Table 2**, the attributes for the weather data includes precipitation/snow and temperature values.

If the weather data we currently intend on using does not provide us with enough information for analysis we have access to data sets that includes information such hourly weather data and wind speeds.

5. EVALUATION METHODS

In order to evaluate our result, we plan on correlating all crime data with different characteristics of the environment. Firstly, our data set provides us with a time, coordinates, and the description of the crime. Using this information we can set up a map of the crime and location of the crime. After this, we will take the time and location in town, and compare it with the temperature. What we can do is create a window of temperatures, and sum up all of the crimes that fall in that window. With this, we can create a ratio between temperature and the amount of crimes. We can then plot the difference between these ratios to analyze the behavior displayed between crime and the environment.

From prior information, as mentioned in **2. LITERATURE SURVEY**, we expect to see some type of average increase in crime as the temperature rises.

6. TOOLS

For us to best find correlations between our data sets we will use a few different tools in order to accurately find patterns. In order to arrange and look at our data we will utilize excel. In excel we can easily clean and arrange our data in whatever way we desire. We can also use excel for some basic analytics. By making use of highlighting certain rows, or placing certain columns next to one another, we can make some initial predictions about possible patterns.

We will be storing our data sets into databases using Microsoft SQL Server. By using SQL server we can create specific queries to look at certain patterns. For example, if I wanted to see weather types for a specific date range and compare that with the crime type and location for that date range, we could easily create a SQL query to give us that data. We could then take that data and put it into excel for further analysis.

The main tools we plan on using for data analysis are WEKA and python. WEKA uses machine learning to help specify patterns within a dataset. WEKA also has a fairly simple user interface that would really help us identify cor-

Table 1: Crime data example entries[2] (Subset of available attributes)

Incident ID	Offense Code	Offense Type	Offense Category	First Occ.	Geo Lon	Geo Lat	Neighborhood ID
20138493	1202	robbery-business	robbery	1/6/13 2:45	-104.9	39.76	northeast-park-hill
2016638673	3512	drug-heroin-possess	drug-alcohol	10/5/16 14:05	-104.98	39.72	speer
2015250914	1102	sex-aslt-rape-pot	sexual-assault	4/25/15 14:30			regis

Table 2: Weather data example entries[4] (Subset of available attributes)

Station Name	Date	Prcp	Snwd	Snow	Tavg	Tmax	Tmin	Tobs
Denver Museum CO US	20170214	0	0	0		48	26	33
Denver Water Department CO US	20170214	0	0	0		48	22	24

relations faster. We also plan to use python scripts in order for us to analyze our data. Due to the immense size of our data sets, utilizing specific python scripts can help us easily iterate through all of the data to help us find specific values and patterns quickly.

For visualizing our data we plan to use Matplotlib. Matplotlib allows for very easy plotting of data, and offers a wide range of different plot styles. Being able to plot the data in many different graph styles could really help us identify outliers, as well as patterns. We also plan to use Jing for any of our screen-grab or screen-casting needs.

7. MILESTONES

To begin with, we will need to clean up our data. Our first milestone will be data cleaning and pruning. It may also be necessary for us to fill in crucial missing data with matching data found from other sources. After that, we will need to arrange and sort our data in a way so as to easily compare each with the others. This will make identifying patterns, locating specific time periods in the data set, and isolating important segments of our data much more efficient and smooth.

Once that is complete, we will be primed to begin our analysis of the data. We will begin by looking for general patterns over a large time period. In doing this we will see which years or decades had the most crime overall, and we will be able to line that up with the weather data to see if there is anything in common. From there, we can go deeper, using smaller time scales to find more precise patterns.

We will document our findings as we go along, and keep our collaborated work updated on our Github. The next milestone will be to find a solid pattern in our data that we can graph or document and submit to Github.

We will continue to adjust and refine how we work as we better understand our data and how it relates, but for now our next milestone will be to answer our Problem Statement questions. We will continue to algorithmically scan through our data to find patterns and correlations until we have reached a point where we can sufficiently answer these questions.

Finally, we will need to create our presentation to show our findings to the class. Our final milestone, for now, will be a clear and concise project presentation. We want to elegantly explain our work to our peers in a way that will intrigue and inform them. By this point, we hope to have a solid understanding of our data and how it all relates so that we can thoroughly answer any questions our classmates may have.

8. PEER REVIEW

After peer review, we have decided to keep with the proposed project. We have chosen to analyze the environment and its effects on crime in the denver-metro area of Colorado. Initially, we decided to analyze only temperature, and location in our results of crime data, however the peer review showed us that possible many other effects of the environment come into play in human nature. However, we plan on keeping our data analysis only to weather and location unless we have extra time to go into more detail about other factors. It was proposed that we would analyze the effects of the moon and other space phenomena on crime. This would be a very interesting thing to cover, but would take a lot of time to find some data about this. As of now, we do not plan on pursuing this information, however if the timing allows, we could.

Hearing about other peoples project really showed us to be specific and careful with the planning of ours. It is far too easy to get overwhelmed by the amount of information and do far too much with it. Our goal is to demonstrate what we want to demonstrate, and display it in a manner that is logical, and easy to understand.

9. REFERENCES

[1] M. Burke, S. Hsiang, and E. Miguel. Weather and violence. *The New York Times*, August 2013.

[2] City and C. of Denver. Crime. https://www.denvergov.org/opendata/dataset/ city-and-county-of-denver-crime.

[3] D. P. Department. Denver crime map. https://www.denvergov.org/content/denvergov/en/ police-department/crime-information/crime-map.html, March 2017.

[4] N. C. F. E. Information. Climate data online. https://www.ncdc.noaa.gov/cdo-web/datasets.

[5] C. P. Thakur and D. Sharma. Full moon and crime. *British Medical Journal*, 289:1789–1791, December 1984.

[6] E. van Zanten. Crime vs. temperature. http://crime.static-eric.com/#top.