Crime Time, Together with Weather

[Progress Report]

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 being 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

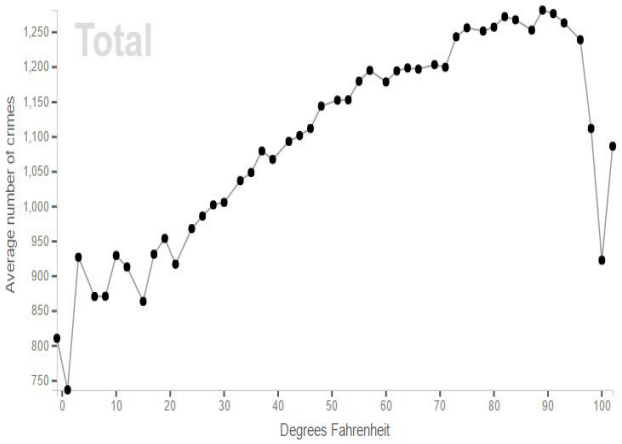

Figure 1: Average number of crimes vs temperature (crime.static-eric.com)[6]

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 preprocess 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 pro-

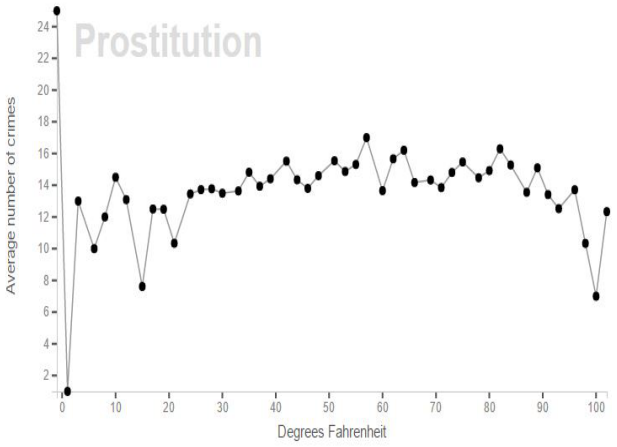

Figure 2: Prostitution crimes vs temperature (crime.static-eric.com)[6]

cess. 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 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 Oceanic and Atmospheric Administration(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 mySQL Server. By using a 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

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

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 correlations 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.

We may also utilize RapidMiner Studio. It is a data mining/analysis tool that offers a lot of visualization and commonly used data analysis methods. It can easily work with a database or ordinary file and compute things like correlation with minimal effort.

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.

7.1 Achieved so far

Most of the work we have completed up to this point has mainly been the preprocessing and organizing of the data for analysis. While this initial work has not resulted in a lot of direct visual results, it will hopefully allow us to more easily complete the difficult and time intensive analysis portion.

The initial crime data was fairly messy and had a lot of missing data and other various issues. There were hundreds of thousands of incidents logged in the crime data set that were not even crimes, in most cases they were traffic accidents because the data includes all incidents reported by the Denver Police Department. Since we are only focusing on crimes, those data points had to be removed.

In terms of the weather data, it needed a lot of work. As expected, there were many errors and inconsistent data values recorded by the hundreds of Denver based weather stations for things like temperature or precipitation. Most of the data was probably automatically recorded daily by data instruments which obviously will make mistakes occasionally. Some of the errors were easily removed/ignored because the data usually included a separate error flag indicating the station data was bad. However, there were many data entries that appeared normal at a glance but in reality were outliers that could severely alter our analysis. In some instances, the entries had values like -9999 indicating no data instead of using the error flags. In other instances, which were much more difficult to detect, was when a data value for something like the maximum daily temperature equaled something like 150 degrees. Obviously to anyone this is clearly a bad value and should be ignored, but when having to work with all the various data types it became increasingly difficult to have to detect these noisy data values. It was not as simple as checking if the value was null, but instead we had to use reasoning involving our real world understanding of the weather to account for any bad data.

Once we completed a lot of the data cleaning we began to integrate the data. For much of the weather data there was multiple station readings for the same day. In most instances, for things like daily max temperature or total pre-

cipitation, we decided to average the values for any data on the same day. This obviously solved the issue of data redundancy in terms of having multiple weather values for a single day, but it also gave us a more accurate data set for the weather near Denver by combining all the individual data sources.

After getting much of the data into a more usable form, we loaded the data into a mysql server. The server will act as our data warehouse where we will hold all the data to be used for analysis. We began creating some data marts that will be used for certain individual aspects of our analytical work. For instance, we took all the crime data and did things like grouping by type of crime or day in order to create tables that only held specific data. This will hopefully allow us to more easily work with the data because we can focus on a subset of it instead of every entry. If we have to do anything computationally intensive it will also potentially help us out, by avoiding data that is not relevant. If we only want to focus on one type of crime we can simply look at the couple thousand reports for that crime instead of having to iterate over every single report.

We also combined much of the weather data and crime data together. For the most part it including combining the date of a crime(s) with the respective weather data for that day. Some precautions had to be made when working with crime data that we aggregated. The primary reason was because we couldn't simply add all the crimes together that occurred at a specific temperature and then say this is the number of crimes that occur at this temperature vs the number of crimes that occur at this different temperature. For obvious reasons, some temperatures may have occurred more often, for instance there were more days that were 70 degrees than 10 degrees. This would obviously make it appear as if crimes occur more at 70 degrees than at 10 degrees, if the number of days was not taken into account. To solve this we simply took the average of crimes per day. This may seem like a simple solution and when looking at all crimes, as mentioned in **8. RESULTS SO FAR**, it worked. However, when working with smaller subsets of specific crimes it turned out be an issue. By having fewer occurrences to work with the data became too variable and didnâĂŹt show any real results.

In terms of actual analysis of the data, we have only begun to work on that aspect. We have created some basic plots, using mostly matplotlib and python. Primarily focusing on the average number of crimes occurring at certain temperatures or other weather events. We have also done some work with crimes occurring at certain times of the day.

7.2 Remains to be done

As stated previously in **7.1 Achieved so far** section, we need to do more digging in terms of our data analysis and related work. With the data cleaned and grouped, we will be able to do the analysis much more efficiently. This should make the analysis portion much quicker, especially with the graphs we have, to easily reference what time of day we will be focusing on for that crime. Of course, we will still need to be critical with our data during our analysis, as there could very well be some factors to consider that we have not thought of. For example, in our theft related data group (theft of motor vehicle, theft of parts from motor vehicle, item theft from vehicle), the reported time of theft may not be entirely accurate. The time window that the crime could have taken place is the largest for this group, so we will need to keep that in mind during our analysis to avoid drawing false correlations. What we have done so far with our data is mostly surface level. Therefore, we still have work to do in terms of mathematical calculations and number crunching. This is very important for what we are looking at with temperature/weather and crime. We obviously want more robust evidence to support the correlations we find, instead of just graphs and observational ideas. We plan to use some algorithms we have learned from class in order to do this.

While we have begun to look at the correlations between weather, crime, and time, we still wish to look at how location plays a factor in crime. In our crime data set we have a column that specifies the address of the incident, as well as the geographic coordinates. By looking at the address and or geographic coordinates of the crime, we can begin to analyze correlations between location and crime. To make this easier we plan to cluster location based on proximity, and then compare it to a cluster of crime data based on crime type. To cluster the data we plan to either partition it using the k-medoids approach, or a model-based approach. Once the data is put into clusters we should be able to find patterns between the two by looking at things like the frequency of certain crimes in specific area.

8. RESULTS SO FAR

Most of our results thus far show a very basic look at our data and analysis. While we hope to look at more than just crimes and their relation to temperature or time of day/year it is what we have to show right now.

The biggest area we planned to focus on was weather, primarily temperature, and crime occurrence. In **Figure 5** our data shows that there is some positive correlation between the average number of daily crimes and increasing temperature. There is some variance in data points, primarily because some temperatures do not occur all that often(really cold or hot temperatures) causing there to be a fewer number of days to average the crime rate over which ultimately causes some variability. This can be seen in the random spikes in the yellow **Avg Crimes** line. However, when the data is fitted with a polynomial regression order 4 line, the dashed red **Trend** line, it is clear there is a correlation between crime and temperature. At around roughly 50 degrees the number of average daily crimes increases up until about 90 degrees where it begins to fall down again. It should be mentioned that the temperatures are the maximum temperature of the day and not the temperature at which a crime occurred. While it is probably safe to assume that most crimes occurred at some temperature within reason to the max temperature, it is possible for instance that a crime occurs at night where the temperature could be significantly lower. When using the minimum daily temperature instead of the max, a very similar graph is generated.

Another area we have looked at and generated some results for has been with crime occurrence and the time of day. In **Figure 3** the percentage of crimes occurring at each hour is graphed. It clearly shows a lower percentage of crimes occur at the early morning and slowly increases throughout the day. There is some random spikes at times like 8am and noon. While correlation does not mean casual, the graph makes basic intuitive sense as we would expect less crimes to occur when most people are asleep. Or likewise, more crimes to occur at a time like noon when more people are

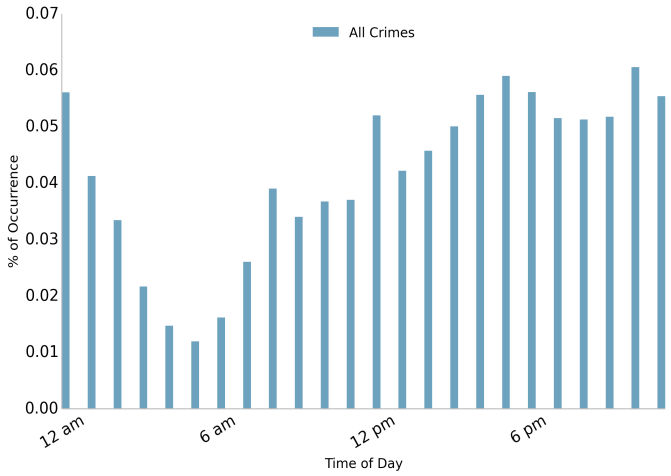

Figure 3: Percentage of all crimes occurring at certain hours of the day

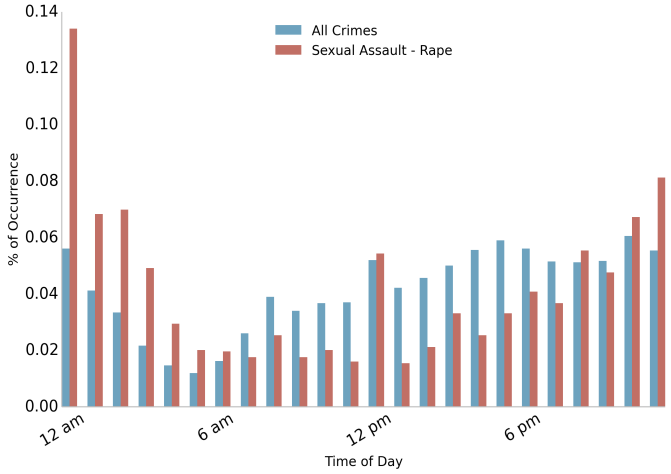

Figure 4: Percentage of sexual assaults(red) and percentage of all crimes(blue) occurring at certain hours of the day

going out for lunch. These are obviously only hypothesis to the results and there could be some completely other reason for the data we see.

We also dove into looking at more specific crimes and how they compare against all the crimes that occur. While the average of all crimes causes what is shown in **Figure 3** where there amount of crimes dies down pretty quickly at around midnight, when looking at sexual assault rape crimes in **Figure 4** something completely different is shown. There is a much higher number of sexual assaults occurring at the early mornings especially when compared to all crimes.

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.

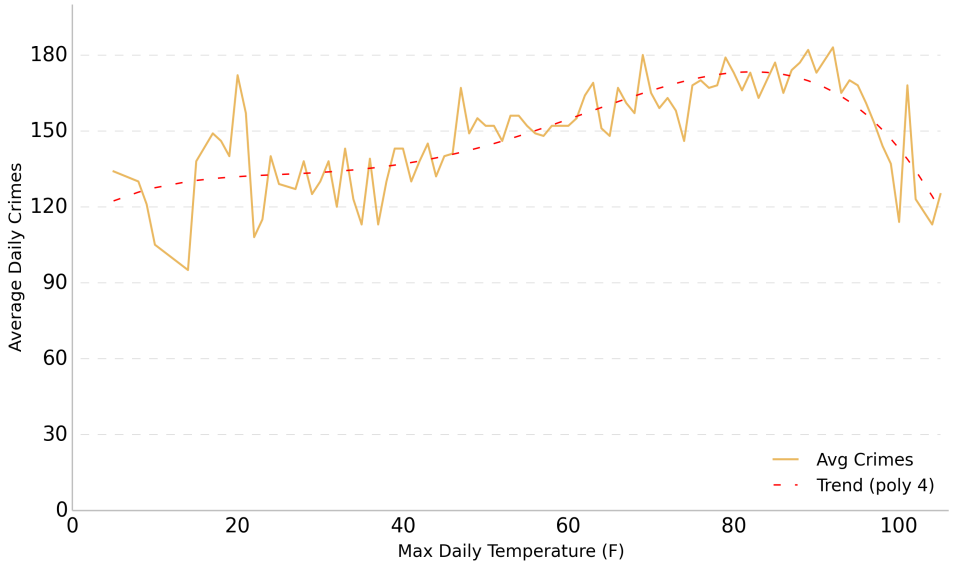

Figure 5: Average number of daily crimes vs the maximum daily temperature

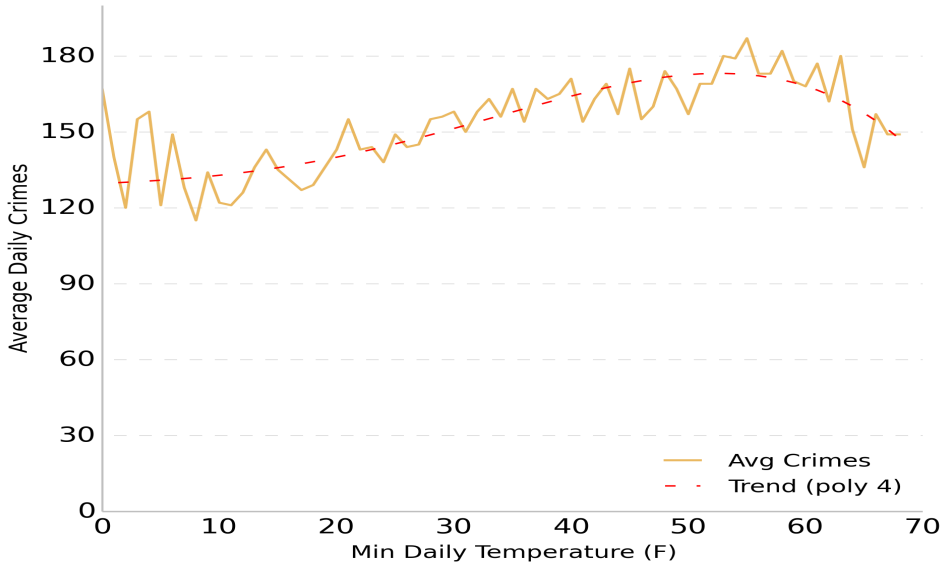

Figure 6: Average number of daily crimes vs the minimum daily temperature