Investigating Factors that Affect Rates of Crime in Milwaukee

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

• ACM copyright: ACM holds the copyright on the work. This is the historical approach.

• License: The author(s) retain copyright, but ACM receives an exclusive publication license.

• Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Verdana 7 point font. Please do not change the size of this text box.

Each submission will be assigned a unique DOI string to be included here.

# Abstract

Marielle Billig

Marquette University

Milwaukee, WI 53233, USA

[marielle.billig@marquette.edu](mailto:marielle.billig@marquette.edu)

James Reilly

Marquette University

Milwaukee, WI 53233, USA

[james.reilly@marquette.edu](mailto:james.reilly@marquette.edu)

Katy Weathington

Marquette University

Milwaukee, WI 53233, USA

[katy.weathington@marquette.edu](mailto:katy.weathington@marquette.edu)

This paper is a preliminary foray into determining if and how various community descriptors can relate to crime rates in the area. The ultimate goal is to develop an interactive tool that allows the user to explore how different factors and parameters can affect the model. The purpose of this iteration is initial analysis and descriptive statistics.

# Author Keywords

Crime; Milwaukee; Regression; Police Districts; Aldermanic Districts; Parks; Communities;

# ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

# Introduction

This format is to be used for submissions that are published in the conference publications. We wish to give this volume a consistent, high-quality appearance. We therefore ask that authors follow some simple guidelines. In essence, you should format your paper exactly like this document. The easiest way to do this is to replace the content with your own material.

# Literature Review

Accepted extended abstracts and papers will be distributed in the Conference Publications. They will also be placed in the ACM Digital Library, where they will remain accessible to thousands of researchers and practitioners worldwide. To view ACM’s copyright and permissions policy, see: <http://www.acm.org/publications/policies/copyright_policy>

# Method

## Data Acquisition

The data for this project is publicly accessible through the city of Milwaukee’s data portal. The crime data is a subset of the data used in previous research. As a byproduct of the previous research, this crime data organized in 120 different files for each year and month.

The dataset including the data about each Police District was compiled manually because the City of Milwaukee’s interface required GUI inputs and outputted a PDF of the resulting data. While it would have been possible to automate this process, considering there were only 7 police districts, it was faster to manually parse the data and create a csv.

## Data Cleaning

Data Cleaning was a surprising large endeavor considering that the data was already very processed. However, the crime data was formatted in .js files that were designed to be easily read and used by javascript web apps. This meant that a significant amount of parsing was required to be able to represent the crime data in python as a pandas DataFrame. Anticipating future needs, special cleaning was applied to the address fields to remove unnecessary characters and enable fuzzy matching. Dates and times were also converted from string to python datetime dates. To reduce the amount of time to load and initialize all the variables in python, a library called pickle was used to save the variables so that they could be loaded directly instead of recalculated in the future.

After both sets of data had been loaded, they had to be merged. In this initial stage, a simple count of all the crimes in each district was calculated. However, future work will likely include breaking this down further by year, month, and or crime type.

## Generating Descriptive Plots

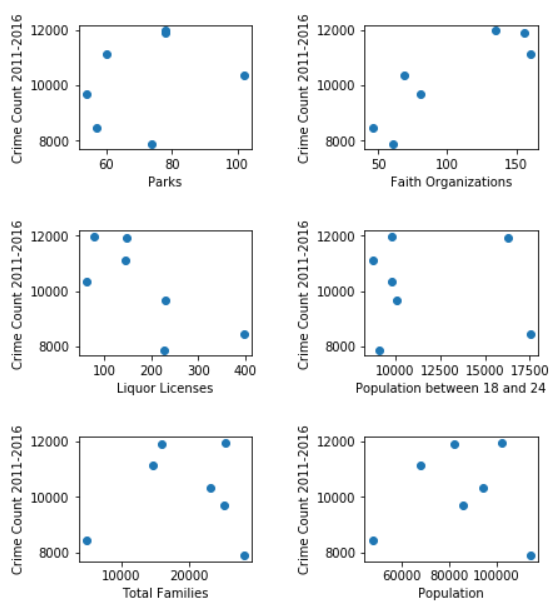
After cleaning the data, generating descriptive statistics was relatively simple. The matplotlib.pyplot library was used to generate simple scatter plots to help gain an initial understanding of the relationships between the different fields.

# Results

Overall there were 71370 records for crimes that were considered. These records were limited to either simple assault arrests, or theft arrests and the date range for the data extended between January 1, 2011 and January 1, 2016.

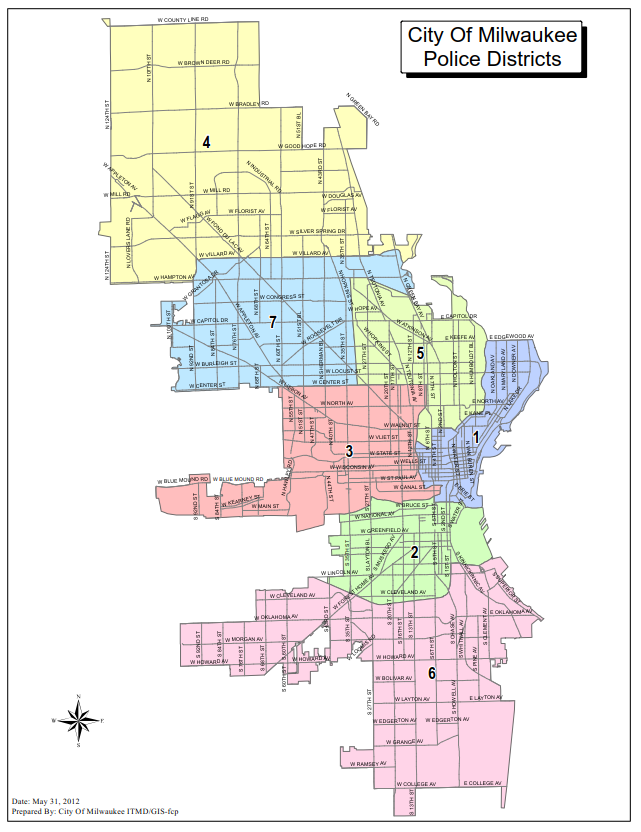
The police district data set was much smaller, with only 7 records, but this will be addressed in the future work section. Relevant fields in the dataset include the population of the district, the area, the distribution of the population ad different ages, the numbers of men and women, the total number of families, the total number of liquor licenses, faith organizations, and parks.

To gain a preliminary understand of the data, the following scatter plots were created to investigate if any of the fields had an obvious relationship with the crime count.



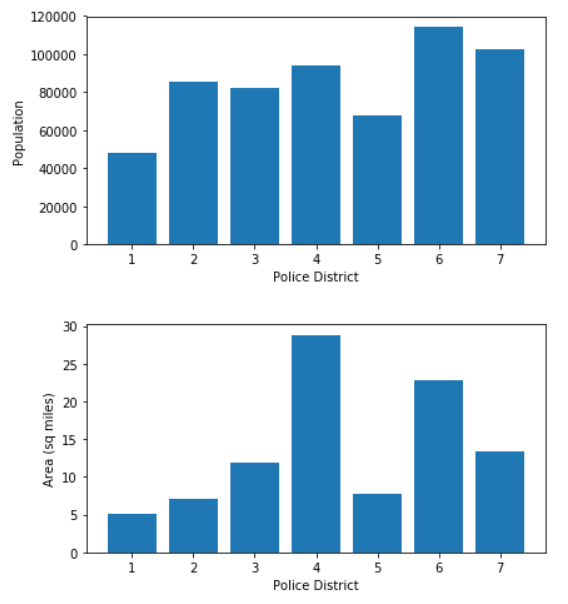
**Figure 1:** Scatter plots showing the relations between several fields and the total count of crimes in each police district

The plots in Figure 1 seem to show that some of the fields appear to have a relationship with the crime count such as faith organizations and liquor licenses. However, a closer look reveals the opposite of what might be expected as Figure 1 shows that an increase in faith organization corresponds to an increase in crime count. This observation led to the realization that the data needed to be normalized by population and area of the police districts since there is a wide variation across the different districts regarding population and area as evidenced the map of the police districts shown below.



**Figure 2:** Police District Map

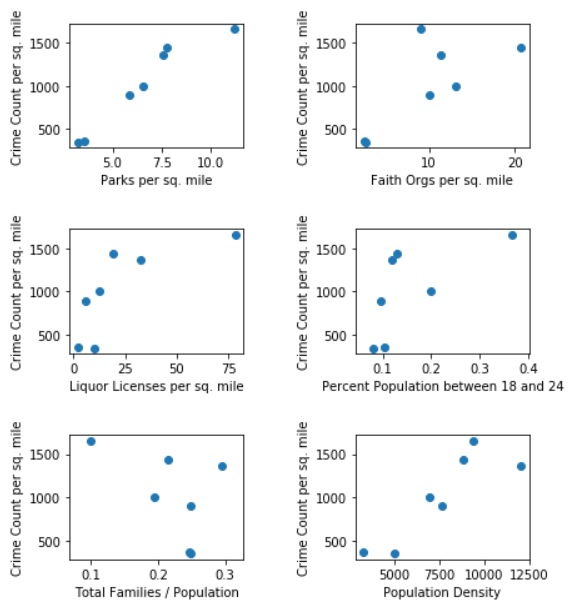
The map clearly shows that some districts, such as District Four, are much larger than others, such as District One.



**Figure 3:** Bar charts showing the distributions of population and area across the different police districts

Figure 3 reiterates the variety in both area and population of the districts. It also points towards the conclusion that is obvious to most city-dwellers that some parts of the city are more densely populated than others.

As a result of these observations, many new fields were calculated. Population, liquor licenses, faith organizations, and parks were divided by the area of the district, and the demographic breakdowns were divided by the total population. Additionally, instead of comparing the fields to the total number of crimes that had occurred in that district, the fields were compared against the crime count divided by the area (Crime density).



These plots show that many of the fields seem to have a relationship with increase crime per square mile, but these plots are just preliminary findings and are not yet conclusive.

# Discussion

TBD

# Future Work

The main area for future work is developing more records than the 7 police districts. Another way of partitioning the city is into Aldermanic districts. These are 15 districts of roughly the same population for which nearly the same data exists, that does for the police districts.

The primary challenge with including aldermanic districts is that the crime data set does not include the aldermanic district directly. To determine the aldermanic district of a crime, the address string has to be compared to the master property list that contains information on all the properties in Milwaukee (also available from City of Milwaukee website). Understandably, this is a very large file, and this would be an expensive operation even if all the strings were perfect matches. Based on initial exploration, approximately half of crime addresses could be matched exactly to an address in the master property list. The other addresses are currently in the process of being matches using python’s difflib fuzzy matching library. This allows for the addresses that are the most similar to the address in question to be returned. A simple mode of the aldermanic districts of these similar addresses is used for the original address.

Beyond letting the aldermanic district calculation finish running, more work also has to be done regarding creating a model that related the field discussed earlier to the crime densities. It would also be interesting to break down the crime counts/densities to subset my month, time or year, to see if the models that work best overall, work best on these subsets as well and vice versa.

Finally, an interactive d3 application must be developed. The current idea it to …. Use sliders….. let users try to beat the model?..... let users choose there input fields and what subset of crime to try to predict?....

# Acknowledgements

Thanks to guha probably, thanks to Milwaukee for making their data public, thanks to Justin and Griffen for making the first steps of getting the crime data

# References

1. @\_CHINOSAUR. 2014. VENUE IS TOO COLD. #BINGO #CHI2016. Tweet. (1 May, 2014). Retrieved February 2, 2014 from https://twitter.com/\_CHINOSAUR/status/461864317415989248
2. ACM. How to Classify Works Using ACM’s Computing Classification System. 2014. Retrieved August 22, 2014 from http://www.acm.org/class/how\_to\_use.html.
3. Ronald E. Anderson. 1992. Social impacts of computing: Codes of professional ethics. *Soc Sci Comput Rev* 10, 2: 453-469.
4. Anna Cavender, Shari Trewin, Vicki Hanson. 2014. Accessible Writing Guide. Retrieved August 22, 2014 from http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/
5. Morton L. Heilig. 1962. Sensorama Simulator, U.S. Patent 3,050,870, Filed January 10, 1961, issued August 28, 1962.
6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal Ubiquitous Comput*. 18, 4 (April 2014), 765-766. http://dx.doi.org/10.1007/s00779-014-0773-4
7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where do web sites come from?: capturing and interacting with design history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02), 1-8. http://doi.acm.org/10.1145/503376.503378
8. Psy. 2012. Gangnam Style. Video. (15 July 2012.). Retrieved August 22, 2014 from https://www.youtube.com/watch?v=9bZkp7q19f0
9. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing.* Indiana University Press, Bloomington, IN.
10. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D Dissertation. Massachusetts Institute of Technology (MIT), Cambridge, MA.
11. Langdon Winner. 1999. Do artifacts have politics? In *The Social Shaping of Technology* (2nd. ed.), Donald MacKenzie and Judy Wajcman (Eds.). Open University Press, Buckingham, UK, 28-40.