DSC520 Final Project

Amie Davis
13 Nov 2019
Crime in Cincinnati

Section 1

- Introduction

  There has been an increase in reported crime in the Cincinnati area. This increase has inflated visibility, and the community is now calling for increased patrols. The city now has a reputation for shootings, and the media has increased their reporting of events. An acceptable solution will deter future crimes and provide a sense of security to the community.

- Research questions
  - Is the increase in visibility due to increased crime or increased reporting?
  - What time of day do crimes occur most often?
  - In what neighborhoods do crimes occur more frequently?
  - Are crimes reported only against certain subjects?
  - Can we predict where crimes are likely to occur?

- Approach

  I plan to address this problem by analyzing Cincinnati crime statistics. I will analyze historical data and build a model to predict times of day and location for similar crimes. Using these predictions, I will recommend changes to patrol numbers and areas to patrol.

- How your approach addresses (fully or partially) the problem.

  Results of this study will show the community and media that patrols are policing all areas, including at-risk areas, providing an added sense of security to the community. It is supposed that an increase in patrolling will deter crimes in those areas.

- Data

  Final Dataset Used:
  PDI (Police Data Initiative) Crime Incidents, City of Cincinnati. https://data.cincinnati-oh.gov/Safety/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf

  - Description: Incidents are the records of reported crimes, collated by an agency for management. Incidents are typically housed in a Records Management System (RMS) that stores agency-wide data about law enforcement operations.

    - Original Created: November 15, 2017
    - Last updated: October 22, 2019, updated daily
    - Rows in Dataset: 380K
    - Columns in Dataset: 40
    - Data Masking: (1) the last two digits of all addresses have been replaced with “XX,” and in cases where there is a single digit street address, the entire address number is replaced with "X"; and (2) Latitude and Longitude have been randomly skewed to represent values within the same block area (but not the exact location) of the incident.

- Required Packages
    - readr
    - ggplot2
    - ggm
    - caTools
    - class
    - ggmap
    - lubridate
    - dplyr
    - gdata
- Plots and Table Needs
    - Plot crime events by date. Compare this year to last year.
    - Histogram showing frequency distribution of crimes by time of day categories.
    - Histogram showing frequency distribution of crimes by neighborhood.
    - Histogram showing reports by demographics
    - Histogram showing subjects by demographics
    - Chart showing results from prediction model.

- Questions for future steps
    - Need to know what variables might be correlated. Will need to plot the data to determine if I see any relationships.
    - May need additional reference data to locate police stations and neighborhoods. (https://data-cagisportal.opendata.arcgis.com/datasets/police-districts)

Section 2

- How to import and clean my data

  See markdown output for data import and cleaning steps.

    - Corrected bad characters in report area field.
    - Corrected incorrect data types for beat, longitude, and latitude columns.
    - Removed bad zip codes.
    - Filtered to Cincinnati zip codes only.

    - Filtered out records with missing geodetic data (lat/long).
    - Grouped records by offense.

- What does the final data set look like?

  See markdown output for structure and header records of cleaned dataset. Additional data available if needed.

- Questions for future steps.

    - Need to convert dates to view by month.
    - Need to determine how best to view geodetic (lat/log) data.

Section 3

- What information is not self-evident?
    - I will plot geodetic (lat/long) data by offense to see if specific offenses are common to certain locations.
    - I will look for a relationship between neighborhood and reporting frequency.
- What are different ways you could look at this data?

  I plan to view the data in different ways by grouping by different variables.

    - Group By Day of the Week
    - Group By Time of Day
    - Group By Month
    - Group By Offense
- How do you plan to slice and dice the data?

  I will split up the date of incident field into day, month, year, hour fields. I will need to review data by various dates.

    - Add a derived time of day – hour field
    - Add a derived alternate to use that ends in September.
    - Filter to use data from the last 2 years.
    - Filter to focus on certain offenses reported in specific neighborhoods.
- How could you summarize your data to answer key questions?
    - Report distribution of reported incidents, by offense, by month, and by neighborhood.
    - Compare data from this year to last year, specifically using date and location data.

- Identify when and when the most commonly reported offenses occur.

- What types of plots and tables will help you to illustrate the findings to your questions?
    - Plot to show crime incidents by month-year. Compare this year to last.
    - Plot to show crime incidents by time of day (hour), considering only the last two years.
    - Plot to show crime incidents by day of the week, considering only the last two years.
    - Plot to show geodetic location by offense, considering only the last two years.
    - Plot to show geodetic location by reported neighborhood, considering only the last two years.

- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

  I plan on creating a kNN model to determine where and when incidents are likely to occur. I am using kNN since I anticipate clusters to form when I plot the data. I also have a known label confirming incident reported. I want the model to be used to predict where incidents may occur, not necessarily where they are reported more often.

- Questions for future steps.
    - I do not know the resolution of reported incidents. There is a field labeled CLSD that may be helpful. However, without knowing convictions, I cannot determine if the incidents reported were true incidents. I can review the CLSD category to determine if any categories are useful.
        - Contacted dataset owner for clarification between “Early Closed” and “Closed” cases.
    - Would like have geodetic information on police station locations to determine how those locations impact the reporting of incidents.

Section 4 – Week 12

I. Introduction:

I sought out to see how I could find insight in crime data from the City of Cincinnati. Different datasets were available for shootings, violent crimes, criminal incidents, and non-criminal reporting. I chose to focus on criminal incidents, since it reflected reporting for multiple offenses. I reviewed geodetic and neighborhood location data to better understand the scope of the problem.

II. Problem Statement:

There has been an increase in crime in the CIncinnati area. This increase has inflated visibility, and the community is calling for increased patrols. The city now has a reputation for shootings, and the media has increased their reporting of events. An acceptable solution will deter future crimes and provide a sense of security in the community.

III. Addressing the Problem:

I sought out to determine whether there was an increase in crime, or just an increase in the reporting of crimes. To do this, I viewed the data in a variety of ways. I viewed distribution of reported incidents and their frequencies. I grouped the data in a variety of ways and derived numeric counts to further analyze. I looked for a correlation between these counts to see if a predictive model could be useful. Most helpful to addressing the problem was the comparison between the number of reported incidents to the closure status of those reports. Did the incidents reported result in arrests or closures? I was able to build a linear model to predict the closure status.

IV. Analysis:

A) Was there indeed an increase in crime? I compared this year's data to last year's data. There was approximately a 3.5% increase in reported incidents since last year. Comparison scatterplots revealed dramatic increases in both December and June. All other comparisons (time of day, type of offense, and closure status) were consistent in both years.

B) Were there particular neighborhoods in which crimes were reported most often? I used scatterplots to look for outliers, and noticed, there was a definite point that stuck out as the number of thefts reported in the Westwood neighborhood. Thefts were reported more than 450% more often than other offenses in Westwood.

C) Is there a particular month or time of day that incidents occur most often? Can parolling be increased during these times? Scatterplots show that incidents are reported most often at midnight, even more so in the summer months. Increased patrolling overnight in the summer is recommended in all neighborhoods.

V. Implications:

The residents of Cincinnati will benefit from the facts drawn out in this report. Rather than being swayed by media reports, the data speaks for itself. The linear models can help determine the closure status of the large number of thefts reported. This can provide focus on where arrests are not being made. I recommend increasing patrolling in areas in which incidents do not result in arrests. Is additional information required from witnesses to result in a conviction? Is there a way the public can help increase arrests in those areas?

VI. Limitations:

I was limited in the demographics available for crime incidents from the City of Cincinnati. Demographics for victims and suspects were largely marked as unknown. I was also constrained by time for this project. There were two additional aspects I wished that I could have analyzed further. I would like to compare incident locations to police stations and satellite stations. I would like to investigate "Closed" and "Early Closed" cases to determine if cases are closed prematurely.

VII. Conclusion:

The analysis has shown there is increased reporting of incidents. Most of these incidents, however, are marked as “Early Closed.” This high number of closure status needs to be addressed. Are they closed because there is not probable cause to arrest? Is the victim not willing to press charges? This analysis opens more questions that require more than a categorical data can provide. However, I believe this analysis pinpoints areas that require further investigation.