Criminal Justice Data Analytics CRIM4002/CRIM6002/SOCI6002

Greg Ridgeway
Rebecca W. Bushnell Professor of Criminology
Chair, Department of Criminology
Professor, Department of Statistics and Data Science

Introduction

This is a data science course. It aims to give students the skills necessary to acquire, organize, link, filter, and visualize datasets. The course is organized around key questions about police shootings, victimization rates, identifying crime hotspots, calculating the cost of crime, and finding out what happens to crime when it rains or a big summer blockbuster is released. However, these questions are designed as motivation for learning the material. I focus less on the questions than on how we arrive at the answers.

The design of the course is to learn about the data sources and tools on the way toward answering questions related to crime and criminal justice. While the main goal is to equip you with the knowledge of data sources and a variety of computational tools, we will work with each data source and tool because we need them in order to answer a key question of interest. Below I list some of the questions, data sources, and tools I intend to cover during the semester.

I do not expect students to have any prior programming experience. In fact, this course is probably not a good choice for students with substantial programming experience. We start on Day 1 computing 2+2 and by the last week we calculate crime counts near schools, transit locations, and street segments.

Graduate students should enroll in CRIM6002/SOCI6002 and will need to complete a final project at the end of the semester. Other than the final project requirement, CRIM4002 and CRIM6002/SOCI6002 are the same course.

Data sources

I include a variety of data sources to give students a broad understanding of the organizations that produce data, the quality of the datasets, and the effort needed to get them in a condition for analysis. These include official government statistics, administrative data, and data scraped off web pages.

- Bureau of Justice Statistics
    - National Crime Victimization Survey (NCVS)
- FBI
    - Law Enforcement Officers Killed and Assaulted (LEOKA)
    - National Incident Based Reporting System (NIBRS)
- City crime and incident data
    - Chicago crime reports
    - PPD officer-involved shootings
- Census Bureau

    - American Community Survey
    - TIGER geography files
- Web scraping
    - Movie revenue by movie by day
    - Temperature and precipitation

Computational methods

We use R (www.r-project.org) for all programming tasks. R is a widely used, free, open source programming language and environment for statistical computing and graphics.

- Basic programming ideas
    - Syntax
    - Data types (variables, numeric, strings, boolean, dates and times, vectors, lists)
    - Flow of control (conditionals, loops, functions)
- Regular expressions (grep, gsub)
- Structured Query Language (SQL), using an R interface to SQLite
- Geocoding, using ArcGIS geocoding service
- Geometry Engine - Open Source (GEOS)
- JSON, to access Census data through APIs
- Parallel processing
- Data quality check techniques
- GPT large language model engine to extract data from text

Course structure

The course is broken down into about seven questions that we aim to answer. On the way to answer these questions we learn about data sources and computational tools. The class time involves me working through a script that demonstrates how to acquire and manipulate the dataset in order to answer the question. The students have their laptops open and work through the steps with me in class. At several points in the script, I give students an exercise, such as redoing what I introduce with a different dataset feature or solving some problematic aspect of the data. These usually take 10 minutes after which I go through the answer (acknowledging that there is rarely one unique approach) and then continue with the script.

Associated with each question is a homework assignment in which the students need to write a script to complete a number of tasks needed to support their answer to the main question.

- What are the trends in police officers killed in the line of duty?
    - Basic R tasks
- Are the number of crimes reported to the police different than the number of crimes that victims report when surveyed?
    - NCVS and UCR data
    - Basic data operations
- Where are hotspots of crime in Chicago?
    - 10 years of Chicago crime data, about 2Gb
    - SQL

    - Mapping and density maps
- What happens to crime when hit movies are released?
    - Webscraping
    - Regular expressions
    - SQL joins
- What happens to crime when it rains?
    - Webscraping
    - Regular expressions
    - SQL joins
- Where do Philadelphia PD officer-involved shootings occur and what happens? When PPD officers shoot someone, to which hospital are they transported?
    - Using ChatGPT to extract data from pdf documents
    - Geocoding
- What is the racial composition of neighborhoods targeted with civil gang injunctions in Los Angeles?
    - Tiger files
    - American community survey
    - GEOS

Grades

Grades are based on 7 to 8 data coding assignments and a final exam.

The homework assignments range from simple data manipulations to more complicated programming exercises. (80% for CRIM4002 students, 60% for CRIM6002/SOCI6002 students)

The final exam is a timed, in-person data analysis, scheduled during the final exam period. Students receive a dataset a few days in advance of the exam and then at the in-person exam answer a collection of basic questions about the dataset using the skills they learned in the course. Students have access to all available tools for the final exam (notes, Google, GPT, etc.) (20% for all students)

CRIM6002/SOCI6002 students must also complete a final project in which they choose a new question, answer it with a new dataset, and use a new tool or technique to answer that question. CRIM6002/SOCI6002 students meet with me in person for 20 minutes for a code review. (20% for CRIM6002/SOCI6002 students)