DATA PREPARATION AND ANALYSIS

PROJECT PROPOSAL AND OUTLINE

Lead and Crime

Analysis of the relation between Lead Contamination in the human and the Crime Rate

Rahul Agasthya
Manasdeep Deb

Nicholas Tating

Ajay Roopesh Mohan
Jay Rodge

Table of Contents

Project Proposal

Abstract

The lead–crime hypothesis is the proposed link between elevated blood lead levels in children and increased rates of crime, delinquency, and recidivism later in life.[1]

Lead is widely understood to be highly toxic to multiple organs of the body, particularly the brain. Individuals exposed to lead at a young age may be more vulnerable to learning disabilities, decreased I.Q., attention deficit hyperactivity disorder, and problems with impulse control, all of which may be negatively impacting decision making and leading to the commission of more crimes as these children reach adulthood, especially violent crimes.[2]

Proponents of the lead–crime hypothesis argue that the removal of lead additives from motor fuel, and the consequent decline in children's lead exposure, explains the fall in crime rates in the United States beginning in the 1990s. This hypothesis also offers an explanation of the earlier rise in crime in the preceding decades as the result of increased lead exposure throughout the mid-20th century.[3]

The lead–crime hypothesis is not mutually exclusive with other explanations of the drop in US crime rates such as the legalized abortion and crime effect. The difficulty in measuring the effect of lead exposure on crime rates lies in separating the effect from other indicators of poverty such as poorer schools, nutrition, and medical care, exposure to other pollutants, and other variables that are predictive of criminal behavior.[4]

[1] Stretesky, Paul B.; Lynch, Michael J. (2004). "The Relationship between Lead and Crime". Journal of Health and Social Behavior. 45 (2): 214–229.

[2] Stewart, W. F.; Schwartz, B. S.; Davatzikos, C.; Shen, D.; Liu, D.; Wu, X.; Todd, A. C.; Shi, W.; Bassett, S. (2006-05-22). "Past adult lead exposure is linked to neurodegeneration measured by brain MRI". Neurology. 66 (10): 1476–1484.

[3] Steel, Daniel (2013). "Mechanisms and Extrapolation in the Abortion-Crime Controversy". In Chao, Hsiang-Ke; et al. (eds.). Mechanism and Causality in Biology and Economics. Springer Science & Business Media. p.188.

[4] Cantor, David; Land, Kenneth C. (1985). "Unemployment and Crime Rates in the Post-World War II United States: A Theoretical and Empirical Analysis". American Sociological Review. 50 (3): 317.

Questions answered

1. Based on counties, how does the overall population of the county affect the lead contamination in the blood? For example, is the concentration of lead higher in the urban people or rural people, which county has the most positive cases, etc.
2. How does the population of a region affect the crime rate? For example, Is the crime rate higher in an urban area or a rural area, etc.

   Note: *For questions 1 and 2, we will be considering the per capita data so that the wide population gap between the counties will not hinder the observations.*
3. Is there a correlation between high crime rates and the high lead contamination in blood?
4. Is it possible to predict if the people of a county were affected by lead contamination using the crime data?
5. Is it possible to predict if a county is vulnerable to a higher crime rate based on the lead contamination data?

Proposed methodology

1. Data Importing and Cleaning
   a. Import the datasets.
   b. Remove unnecessary or redundant attributes.

      Certain columns will not be required. This includes data like the number of cases reported for each crime.
   c. Address null values.

      There are certain cases where the number of tests for lead contamination was not conducted in a particular region of a county. Hence, such regions may have to be removed to ensure an accurate prediction model.
   d. Apply appropriate data transformations
   e. Join data based on location (likely County)
2. Exploratory Data Analysis
   a. Establish a relationship between the population of a region, and the per capita crime rate and the per capita lead contamination.
   b. Explore relationships between lead levels and crime.
   c. Visualize correlations between lead levels and crime rates based on counties.

      Use correlation matrices, scatterplots, etc. to visualize correlation
   d. Attempt to discover hidden insights.
3. Modelling
   a. Linear Regression to predict crime rate.
   b. Classification Models to classify a county's level of risk of increased crime based on lead contamination.

Metric for measure analysis result

- Crime Rate - total violent crimes per capita
    - We aim to establish a relationship between blood lead levels and crime rates with the intention of predicting crime rates amongst counties.
    - Computed by total violent crimes per 100,000 residents using total violent crimes from crime dataset and population from population dataset.
    - Model predictor measured by RMSE
- Accuracy of County Crime Rate Vulnerability Classification
    - We aim to classify a county's vulnerability to increased crime rates based on lead contamination.
    - We measure its accuracy using (true positives + true negatives) / (true positives + false positives + true negatives + false negatives)

Project Outline

Reference Data

http://www.dph.illinois.gov/sites/default/files/publications/lead-surveillance-report-2016.pdf

Data Sets and Sources

Dataset for the Crimes in New York(grouped by county):

https://data.ny.gov/Public-Safety/Index-Crimes-by-County-and-Agency-Beginning-1990/ca8h-8gjq/data

Number of Columns: 15
Column Characteristic: Multivariate
Number of Observations(rows): 19,956
Missing Values: Yes

Column Name	Datatype	Description
County	String	The name of the county
Agency	String	The name of the reporting security agency.
Year	Numeric	Year in which the crimes were reported.
Months Reported	Numeric	Number of months crimes were reported.
Index Total	Numeric	Total number of all kinds of crimes reported.
Violent Total	Numeric	Number of violent cases reported.
Murder	Numeric	Number of murders reported.
Rape	Numeric	Number of rapes reported.

Robbery	Numeric	Number of robberies reported.
Aggravated Assault	Numeric	Number of aggravated assaults reported.
Property Total	Numeric	Total number of property thefts reported.
Burglary	Numeric	Number of burglaries reported.
Larceny	Numeric	Number of larcenies reported.
Motor Vehicle Theft	Numeric	Number of motor vehicle thefts reported.
Region	String	Whether the region is New York City or not.

Dataset for childhood Blood Lead Contamination:

https://health.data.ny.gov/Health/Childhood-Blood-Lead-Testing-and-Elevated-Incidenc/d54z-enu8/data

Number of Columns: 14
Column Characteristic: Multivariate
Number of Observations(rows): 29,807
Missing Values: Yes

Column Name	**Datatype**	**Description**
County	String	County of residence
County Code	Numeric	County FIPS code
Year	Numeric	Year of lead test data collection
Zip	String	Zip code of residence

Tests	Numeric	Number of children tested in that zip code
Less than 5 mcg/dL	Numeric	Number of children identified to have less than 5mcg/dL lead concentration.
5-10 mcg/dL	Numeric	Number of children identified to have 5-10 mcg/dL lead concentration.
10-15 mcg/dL	Numeric	Number of children identified to have 10-15 mcg/dL lead concentration.
15+ mcg/dL	Numeric	Number of children identified to have more than 15mcg/dL lead concentration.
Total Elevated Blood Levels	Numeric	Total number of children confirmed to have elevated blood lead level (>10mcg/dL is considered elevated).
Percent	Numeric	Percent of children found to have elevated levels out of total children tested.
Rate per 1000	Numeric	Rate per thousand of children found to have elevated levels.
Zip Code Location	Location	Latitude and Longitude of Zip code.
County Location	Location	Latitude and Longitude of the County.

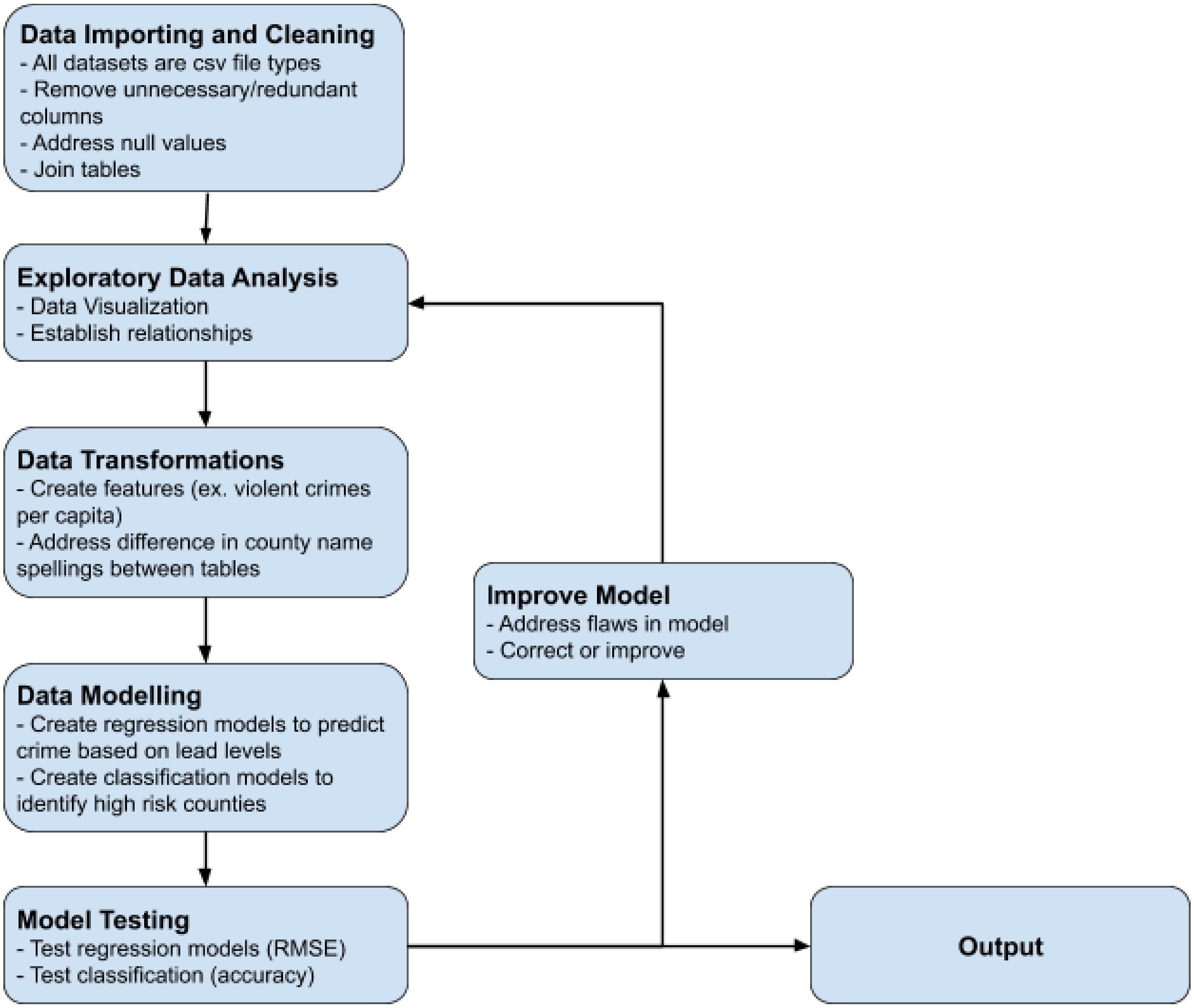
Dataset for Populations of New York Counties:

https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k/data

Number of columns: 5
Column Characteristic: Multivariate
Number of Observations (rows): 3,402
Missing values: No

Column Name	**Datatype**	**Description**
FIPS Code	String	Federal Information Processing Standards codes that identify each geographic area.
Geography	String	Geographic area name
Year	Numeric	Year for which the population is calculated
Program Type	String	Type of Census (Census Base Population, Intercensal, Postcensal)
Population	Numeric	Number of residents

Data processing and pipeline


Data Importing and Cleaning
- All datasets are csv file types
- Remove unnecessary/redundant columns
- Address null values
- Join tables
Exploratory Data Analysis
- Data Visualization
- Establish relationships
Data Transformations
- Create features (ex. violent crimes per capita)
- Address difference in county name spellings between tables
Data Modelling
- Create regression models to predict crime based on lead levels
- Create classification models to identify high risk counties
Model Testing
- Test regression models (RMSE)
- Test classification (accuracy)
Improve Model
- Address flaws in model
- Correct or improve
Output

The above figure denotes a general idea/flow of the Data processing and pipeline for the Model Building.

The data is already in the CSV format, and therefore no need to convert it to other file type.

Both the datasets mainly have numeric as the datatype in most of there necessary columns.

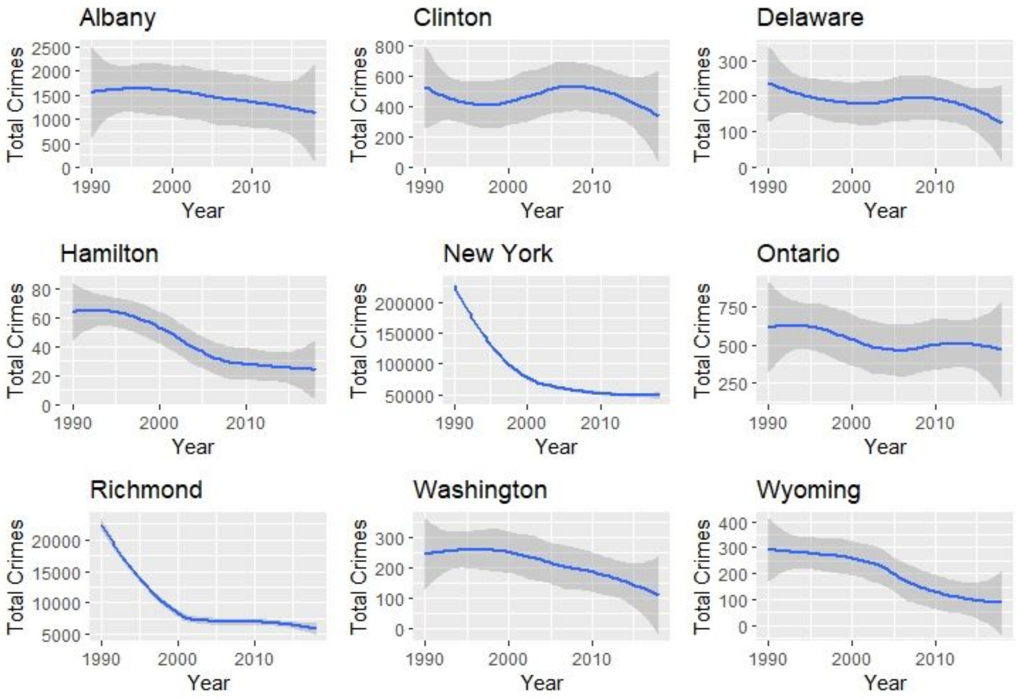
The columns which are String in data type (such as County, Region in Crimes dataset)will be converted into categories by using label encoder which converts this kind of categorical text data into model understandable numerical data.

The dataset also contains missing values or NaN which needs to be addressed before performing any operation. There are two approaches for handling missing values: imputation and removing NaN values.

SInce most of the data has NaN values, removing rows with NaN values would affect the distribution of the data. Hence, imputation has to be carried out which will replace NaN values with mean of the column, which can be done by simple.impute in R.

Data stylized facts

Figure: Crime rate in some counties over 18 years.



Model selection

To examine the association between blood lead levels and the crime rate, we will use a multivariate linear regression model to find a slope factor relating crime rate in a county to the average blood lead level in the county's population. As we make progress in the project, it is intended that we make the prediction using multiple models (like k-nearest-neighbors, Naïve Bayes, etc), and give a comparison of the accuracy of various models.

Tools

- Software/Applications
  - RStudio
  - R
- Libraries
  - tidyverse
  - ggplot2
  - dplyr
  - knitr
  - corrplot

- Project Management and Source Control
    - GitHub
      For maintaining the code and the data.
    - Slack
      For team discussions and primary means of communication.
    - Trello
      Assigning tasks for each member of the team.