Impact of City Investment on Violent Crime

Team 51 MGT 6203 Group Project Proposal

TEAM INFORMATION

Team #: 51

Jenny Chen (jchen3176)

- Data Scientist at a healthcare startup with background in strategy and actuarial.
- Majored in math and econ at UVA.
- Have experience with churn modeling and healthcare analytics.

Jonathan Cascioli (jcascioli3)

- Data Analyst at a corporate construction company.
- Background in construction engineering.
- Undergraduate degree in Construction Science and Management.
- Perform analysis tasks daily for the various business sectors at my company.
- Thus far mostly high-level exploratory or low level statistical calculations.

Makrand Kalyankar (mkalyankar3)

- Data Engineer at a retail ecommerce company
- Educational background in Mechanical Engineering and Physics
- Professional experience in all roles in data engineering.
- Have taken most of the required courses on OMSA

Meir Weiner (mweiner30)

- Battery engineer / lab manager at a startup focusing on battery data processing and analysis. Majored in chemical engineering.

Xiaolu Su (xsu73)

- Junior architect at an AE firm
- Graduated with B.S. in Architecture
- Completed half of OMSA and worked on some solo projects on smart-home and game simulation.

OBJECTIVE/PROBLEM

Project Title:

Background Information on chosen project topic:

The United States is diverse across geographies in many aspects, including the **rate at which violent crimes are committed**. There are many known factors that contribute to crime instances within a city, including demographic makeup (ex: race, gender, age, income), cultural factors (religion, general citizen values), and even climate and degree of urbanization.

Primary Research Question (RQ):

What policies can a local government change to make communities safer and reduce the instances of violent crime?

For the purposes of this project, we wanted to focus on only the factors that a city/ town governing entity is able to influence in an ethical way. Broadly, these include:

1. Law enforcement (ex: police officers, police stations, security cameras)
2. Social programs (ex: YMCA's, Planned Parenthood, homeless shelters, food banks)
3. Education (ex: after-school programs, school budgeting & staffing, public training)
4. City Planning (modes of transportation, streetlamp installation, house vacancy rates)
5. Employment (ex: target industries for job creation, job finding assistance)

Technical Problem Statement:

The purpose of our analysis and investigation is:

1. [inference]: determine which factors are most influential when it comes to violent crime occurrences
2. [prediction]: if any of the influential factors are changed for a city, how can they expect their crime rate to increase/ decrease?

Possible Supporting Research Questions:

1. What are factors that correlate with crime rate and to what degree are they related? What are the most significant variables?
2. How can a city government act on those factors in an ethical and responsible way, with limited budget/ resources?

Business Justification:

Above all, it should be the priority of every local government to ensure the safety of its citizens. Therefore, it is expected that a portion of taxpayer dollars should be spent in support of this goal. This analysis will help local governments decide how to allocate their limited resources to achieve a tangible, metric-driven result, and communicate findings and predictions transparently.

Additionally, the inference driven by our models can help pinpoint the factors that most influence crime rates, which has business value even at the citizen-level.

DATASET/PLAN FOR DATA

Data Sources (links, attachments, etc.):

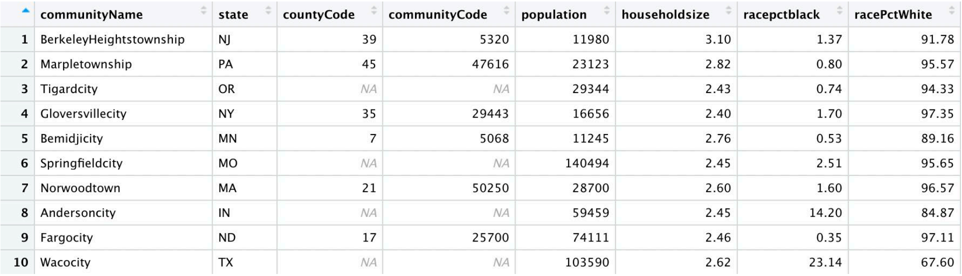
1. [core data source] Crimes in US Communities: https://www.kaggle.com/datasets/michaelbryantds/crimedata
   a. Data schema (for most fields): https://archive.ics.uci.edu/ml/datasets/communities+and+crime
2. City
   a. Web scraped City level data for additional features: https://gatech.box.com/s/saetrysw34gghaf6yysky5nngd6ifq31
3. Zip Code
   a. https://www.unitedstateszipcodes.org/zip-code-database/
4. Sources Under Review for Project compatibility/need:
   a. School Finance Data
      i. https://www.census.gov/programs-surveys/school-finances/data/tables.html
   b. National Center for Education Statistics
      i. https://nces.ed.gov/edfin/search/search_intro.asp
   c. FBI Crime Data Explorer – National incident based reporting system (NIRBS)
      - https://crime-data-explorer.app.cloud.gov/pages/downloads#nibrs-downloads

Data Description

Crimes in US Communities:

Our core data source comes from a UCI repository. It is actual historical data from the 1990's on crime and related factors in US communities, joined from a combination of: 1990 census data, 1990 law enforcement survey data, and 1995 FBI crime occurrence data.

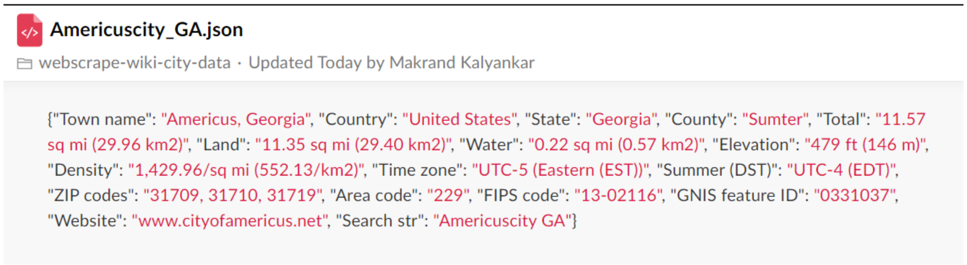
The primary key for this table is the combination of the communityName and state columns which make up a unique "community" identifier.



	communityName	state	countyCode	communityCode	population	householdsize	racepctblack	racePctWhite
1	BerkeleyHeightstownship	NJ	39	5320	11980	3.10	1.37	91.78
2	Marpletownship	PA	45	47616	23123	2.82	0.80	95.57
3	Tigardcity	OR	NA	NA	29344	2.43	0.74	94.33
4	Gloversvillecity	NY	35	29443	16656	2.40	1.70	97.35
5	Bemidjicity	MN	7	5068	11245	2.76	0.53	89.16
6	Springfieldcity	MO	NA	NA	140494	2.45	2.51	95.65
7	Norwoodtown	MA	21	50250	28700	2.60	1.60	96.57
8	Andersoncity	IN	NA	NA	59459	2.45	14.20	84.87
9	Fargocity	ND	17	25700	74111	2.46	0.35	97.11
10	Wacocity	TX	NA	NA	103590	2.62	23.14	67.60

There are over a hundred columns in this source, but not all are feasible for use. We will limit our analysis to only the features that a local government can influence, stripping out most demographic and wealth-related columns.

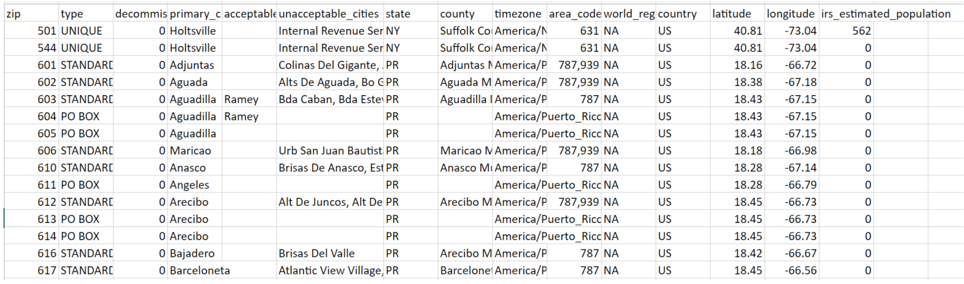
Makrand has done some preliminary web-scraping, providing us with data on city/ county information.



Americuscity_GA.json

webscrape-wiki-city-data · Updated Today by Makrand Kalyankar

{"Town name": "Americus, Georgia", "Country": "United States", "State": "Georgia", "County": "Sumter", "Total": "11.57 sq mi (29.96 km2)", "Land": "11.35 sq mi (29.40 km2)", "Water": "0.22 sq mi (0.57 km2)", "Elevation": "479 ft (146 m)", "Density": "1,429.96/sq mi (552.13/km2)", "Time zone": "UTC-5 (Eastern (EST))", "Summer (DST)": "UTC-4 (EDT)", "ZIP codes": "31709, 31710, 31719", "Area code": "229", "FIPS code": "13-02116", "GNIS feature ID": "0331037", "Website": "www.cityofamericus.net", "Search str": "Americuscity GA"}



zip	type	decommis	primary_c	acceptable	unacceptable_cities	state	county	timezone	area_code	world_reg	country	latitude	longitude	irs_estimated_population
501	UNIQUE	0	Holtsville		Internal Revenue Ser	NY	Suffolk Co	America/N	631	NA	US	40.81	-73.04	562
544	UNIQUE	0	Holtsville		Internal Revenue Ser	NY	Suffolk Co	America/N	631	NA	US	40.81	-73.04	0
601	STANDARD	0	Adjuntas		Colinas Del Gigante, .	PR	Adjuntas M	America/P	787,939	NA	US	18.16	-66.72	0
602	STANDARD	0	Aguada		Alts De Aguada, Bo G	PR	Aguada M	America/P	787,939	NA	US	18.38	-67.18	0
603	STANDARD	0	Aguadilla	Ramey	Bda Caban, Bda Este	PR	Aguadilla I	America/P	787	NA	US	18.43	-67.15	0
604	PO BOX	0	Aguadilla	Ramey		PR		America/Puerto_Ricc		NA	US	18.43	-67.15	0
605	PO BOX	0	Aguadilla			PR		America/Puerto_Ricc		NA	US	18.43	-67.15	0
606	STANDARD	0	Maricao		Urb San Juan Bautist	PR	Maricao M	America/P	787,939	NA	US	18.18	-66.98	0
610	STANDARD	0	Anasco		Brisas De Anasco, Est	PR	Anasco Mu	America/P	787	NA	US	18.28	-67.14	0
611	PO BOX	0	Angeles			PR		America/Puerto_Ricc		NA	US	18.28	-66.79	0
612	STANDARD	0	Arecibo		Alt De Juncos, Alt De	PR	Arecibo M	America/P	787,939	NA	US	18.45	-66.73	0
613	PO BOX	0	Arecibo			PR		America/Puerto_Ricc		NA	US	18.45	-66.73	0
614	PO BOX	0	Arecibo			PR		America/Puerto_Ricc		NA	US	18.45	-66.73	0
616	STANDARD	0	Bajadero		Brisas Del Valle	PR	Arecibo M	America/P	787	NA	US	18.42	-66.67	0
617	STANDARD	0	Barceloneta		Atlantic View Village,	PR	Barcelonet	America/P	787	NA	US	18.45	-66.56	0

With the help of Meir, the primary key column of Crimes DataSource was cleaned and has successfully been joined to county/zip code data using a City-State combination. This marks a key step, as having zip codes readily available will allow us to join and aggregate data from a wide variety of government and public sources. As we further develop our analysis, we will continue to filter and gather data related to law enforcement, education, and social spending.

Key Variables:

The **dependent variable** is rate of violent crime. We will likely need to normalize the number of crime occurrences by population for each observation.

The **independent variables** include (but are not limited to) the following: (all non-percentage features will likely be normalized by population).

Law enforcement variables:

- Number of police officers
- Number of police cars
- Police average overtime worked
- Police operating budget

Homelessness Variables:

- Number of people in homeless shelters
- Number of homeless people counted in the street

Social Program Variables:

- Number of social organizations such as YMCA’s, Planned Parenthood, food banks, etc.

Education variables:

- percentage of people 25 and over with less than a 9th grade education
- percentage of people 25 and over that are not high school graduates
- percentage of people 25 and over with a bachelor's degree or higher education
- percent of people who speak only English
- percent of people who do not speak English well

City planning variables:

- percent of housing occupied
- percent of vacant housing that is boarded up
- percent of vacant housing that has been vacant more than 6 months
- percent of people using public transit for commuting

Employment variables:

- percentage of people 16 and over, in the labor force, and unemployed
- percentage of people 16 and over who are employed
- percentage of people 16 and over who are employed in manufacturing
- percentage of people 16 and over who are employed in professional services
- percentage of people 16 and over who are employed in management or professional occupations

The variables we currently hypothesize to be **most important** are:

- Education (specifically, % of people that are not high school graduates)
- Number of social organizations
- Number of police officers.

APPROACH/METHODOLOGY

We will likely attempt 2 types of regression models: a multivariate linear regression and a tree-based model. Although prediction of crime rates is part of our goal, the most important outcome of this analysis is pinpointing what features are most significant (where governments should allocate resources to). Therefore, we are choosing models that prioritize interpretability over prediction accuracy.

Our team has already done a significant amount of pre-processing and data cleaning. We will continue this and start the necessary feature engineering before any modeling is attempted. The core dataset has missing values and needs to be supplemented by ancillary sources, which require some manipulation before joining. Additionally, some of the raw features require slight transformations to be standardized against a community's population size.

We will split our data into training and test sets and compare the models by looking at accuracy over the test set.

To train our CART hyperparameters we will use the rpart package in R* and start with default hyperparameters. We may choose to manually adjust these based on what we deem appropriate for use cases of the data (ex: maxdepth, minsplit). To optimize the hyperparameters, we will use the mlr or caret package in R* and apply k-fold cross-validation, and finally evaluate the results using elbow plots.

**Or comparable libraries in python*

Anticipated Conclusions/Hypothesis:

We hypothesize the following variables with have a positive relationship, I.e. as the independent variable increases the Rate of crime will increase:

- Homeless Variables
- City Planning Variables
- 25+ Has Not graduated high school
- 25+ less than 9th Grade Education
- Non-English Speaking / Poor English Skills

We hypothesize the following variables will have an inverse relationship, I.e. as the independent variable increases crime will decrease:

- All law enforcement variables
- Social Program Variables
- Employment Rate
- 25+ Bachelor's Degree
- Speak Only English

Measuring Outcomes:

- Since our project is intended to inform governmental policy, we seek to align it with the standards of the Census Bureau where applicable and realistic given the timeline and scope of our project.
    - https://www.census.gov/about/policies/quality/standards/standarde1.html
    - https://www.census.gov/about/policies/quality/standards/standardd2.html
- We will filter significant relationships using a P value seeking at least a 90% confidence interval, though 95% would be preferential. If a variable does not exhibit this we will deem it as not contributing value and remove it from the model as well as our findings as having influential impact on policy (inference).

- For our model, we would consider it a success in prediction if the adjusted R squared is 85%. Given the breadth of the problem even 85% might be a high goal. We will continue to refine this as our research continues.

Impact

As mentioned before, understanding the actual relationship between our various variables and the crime rate is what is important. Our emphasis lies on this comprehension because it enables greater insight into the stances one should take on matters of public/governmental policy to have the greatest impact on decreasing crime. For example, if it is shown that greater homeless rates lead to increased crime, a leader in public policy might push for efforts to decrease the homeless population such as through stricter laws on loitering, eliminating laws that allow for public "camping" within city limits, or increasing funds for social programs to educate/employ the homeless. In other words, our model seeks to uncover these relationships to understand the direction one should take on these important issues but does not focus on the exact tool to use.

It is important to note that establishing direct causation of our target variables is beyond the analysis of this project. The intent is to establish a foundation, or proof of concept, on which an extensive investigation could be grown from.

PROJECT TIMELINE/PLANNING

Project Timeline/Mention key dates you hope to achieve certain milestones by:

10/15: data preprocessing: finalize data cleaning and merging, preliminary observations of response variable and key explanatory variables.

10/22: train linear regression and CART models

10/29: film plan presentation video & complete progress report

11/5: CART hyperparameter optimization

11/12: film final presentation video and complete report

Appendix:

Resources:

- Web scrapper to download city information from Wikipedia to enhance the city level data to create features to answer questions being laid out in this proposal. Git-hub Webscrapper-wiki.

Research:

1. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports
    - This paper seems to focus on the effects of education length and attendance to crime rates

   - https://eml.berkeley.edu/~moretti/lm46.pdf

2. More school funding less crime
   - https://www.annenberginstitute.org/events/more-school-funding-less-crime#:~:text=We%20find%20that%20students%20exposed,being%20arrested%20as%20an%20adult.
3. The Intergenerational Effects of Education on Delinquency
   - Speaks to the long term effects and other dynamics
   - http://econ.hunter.cuny.edu/wp-content/uploads/2017/01/Chalfin_Deza_compulsory2016.pdf