Exploring the Correlation Between Regional Urban Government Spending and Crime Rate Data

Ashley Jean
Towson University
ajean4@students.towson.edu

Zakina McGee
Towson University
zmcgee2@students.towson.edu

Mark Shanks
Towson University
mshank1@students.towson.edu

Abstract - Our objective is to find and/or examine the correlation between the crime rate amongst New York City geographical regions and the funding received by those regions from the city. More specifically, we are interested in examining whether or not an increase in funding translates to lower crime rates, and if so, what useful information can be inferred from that model by government regulators. While a majority of prior work includes average household income as a key influencing attribute affecting crime statistics in a given region, prior work exists [1] that explores correlations in government spending and local crime statistics. However, such analyses were performed on high-level federal census data rather than specific city reporting data which may harbor new insights. Using a data model trained on finer-grained data records and new data features, our team intends to draw further predictions on the relationship between funding and crime rate for a given region.

Keywords - Urban Crime, Crime Causality, Government Spending, School Spending, Correlations, Crime Predictive Analytics

I. INTRODUCTION

It is generally known that there is a fundamental correlation between the economic health of a populated region; such as a neighborhood, town, city, and the safety amenity economic externality (crime rate) in that region. Typically, regions with a greater number of higher-income households tend to have lower crime

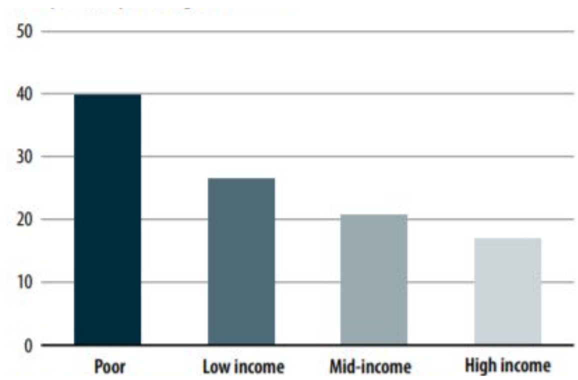


Figure 1: U.S. national violent crime rate with respect to household income level, 2008-2012 (Source:https://www.bjs.gov/content/pub/pdf/hpnvv0812.pdf)

However, less is known about the relationship between government spending and crime statistics. For example, if the government spends money directly on efforts such as blight removal, school facilities, and infrastructure, or indirectly such as by de-regulating rent control policies to initiate gentrification or authorizing construction permits, the residents living in that region will experience an increased standard of living which is generally accompanied by a corresponding reduction in crime[2].

In this project, our team seeks to analyze the impact of government spending in a particular geographic region on the crime rate in that region. In this analysis, each member will explore the predictive power of one of the following three target variables:

1. Spending location (geographic impacts of funds)
2. Spending rate (temporal concentration of funds)
3. Spending type (types of projects funded)

A decision-tree model traine on publically available datasets will be developed using the R programming language, and the performance of each model will be evaluated and compared. A schedule of group work is provided in Appendix A.

II. DESCRIPTION OF THE DATASET

Initially, our team was looking to use data from the Open Baltimore data set website, but this data was found to be sparse and unreliable. Instead, the dataset we've chosen for this analysis comes from New York City's open dataset website, which will still provide a useful means of exploring the relationship between crime rate and government regional spending. Specifically, we pulled two data sets that record crime reports and government school system spending by geographic region.

The city open datasets are freely available to everyone to use without restrictions. We choose these datasets for our analysis because they provided the most accurate information and data needed for our predictions. In fact, the datasets are maintained and updated periodically to ensure accuracy.

The first data set is "Capital Project Schedules and Budgets", which is a listing of government spending on schools indexed by city district from 2003 to 2019, with 10,797 records and 15 features. Initial exploratory data analysis revealed that many records indicate the planned projects have not started yet, so the total useful record count for this analysis is only 5,930 records.

The second data set is "NYPD Complaint Data Historic", which contains 6.5M records of crime reports and 35 features. Our team will likely focus on more violent crimes that have a more severe impact on the perception of public safety within a region. This will help reduce computational complexity.

After cleaning the data, our team will join these two datasets using the "merge" feature in the R programming language. The sets will be joined by the common geographical region.

Unfortunately, for the geographical features, the funding records use "voting district number" whereas the crime records use "police precinct number", and these two regions are not the same. However, in prior research, A. Golub et.al. generated and approximate mapping between New York City precincts and districts *(see Appendix B)*, and this mapping will be applied to our data sets to facilitate joining[3].

III. PRELIMINARY REVIEW

A preliminary review of literature was conducted on recent crime analysis. Grawert, Kimble, and Onyekwere conducted a preliminary review of crime based on available crime data from police departments. When collecting data, the authors searched first for reliable, regularly updated data sources. Accordingly, weekly CompStat reports, monthly reports, and quarterly reports were used wherever possible. Reliable open data portals, such as those maintained by Baltimore and Chicago, were also prioritized. Based on experience indicating that data portals are often slow to update, where data was collected from such sites, estimates were based on crime data as of the end of the third quarter of the year[4].

Where data was collected from CompStat or similar reports, the authors used the most recent report available at the time of analysis. Offense data were categorized according to UCR definitions: Violent crime includes murder, robbery, and aggravated assault. Property crime includes burglary, larceny-theft, and motor vehicle theft. Murder includes murder and non-negligent manslaughter. Overall crime includes all the above. For the final 2018 analysis, the city of Baltimore had a 6,377.6 crime rate per 100,000 and a 1,815.6 violent crime rate per 100,000.

A preliminary review was conducted on literature from Uniform Crime Reporting. The Uniform Crime Reporting Program (UCR) primary objective is to generate reliable information for use in law enforcement administration, operation, and management; over the years, however, the data have become one of the country's leading social indicators [5]. The program has been the starting place for law enforcement executives, students of criminal justice, researchers, members of the media, and the public at large seeking information on crime in the nation. According to UCR, the latest release of crime data in 2018:

City	Population	Violent crime	Murder and nonnegligent manslaughter	Rape[1]	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft	Arson[2]
Aberdeen	16,210	116	0	8	24	84	314	48	256	10	0
Annapolis	39,461	215	1	27	47	140	910	133	722	55	3
Baltimore	605,436	11,100	309	361	5,066	5,364	27,217	6,048	16,794	4,375	138

Figure 2: 2018 Baltimore UCR Crime Data (Source: https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018)

The Authors predicted crime by gathering crime data from directly from city sources — in many cases CompStat reports. To ensure an accurate comparison, the authors accounted for historic variations between UCR and CompStat data in the following manner. Using an equation similar to the one used in our preliminary and updated reports, the authors assumed that the ratio of final, UCR-reported offenses to CompStat offenses would not vary year-to-year.

$$\text{UCR2018(est.)} / \text{CompStat2018} = \text{UCR2017} / \text{CompStat2017}$$

This ratio allowed the authors to "solve" for an estimate of final, UCR-reported offenses based on CompStat figures from both years and UCR data from 2017. UCR2018(est.) = CompStat2018 × (UCR2017/CompStat2017) [5]. For example, if Baltimore reported 200 aggravated assaults to the FBI in 2017, but the 2017 UCR showed that Baltimore experienced only 100 aggravated assaults that year. These data points would suggest that only 100 of the 200 incidents reported by Baltimore police to the FBI met the agency's definition of "aggravated assault." Therefore, in 2018, if Baltimore's publicly available data showed 150 aggravated assaults, this report's methodology would apply the same ratio and assume that only 75 of them would meet the FBI definition of the crime.

IV. PROPOSED METHODOLOGY

Our team will be using the R programming language and associated libraries to perform the proposed analysis on the listed datasets. As a team, our first step will be to perform feature reduction, to eliminate columns that are not of interest in each data set. For example, in the NYC crime data, our current approach is to map each record to a political district using a previously identified precinct-to-district mapping. Since this effort provides the geo-spatial feature for each record, there is no need for the latitude and longitude coordinates.

Next, we will clean the data to remove erroneous, missing or unnecessarily redundant data. Following the cleaning evolution, we will "join" multiple datasets by the geo-spatial feature. Our team will work together to decide what temporal increment will yield the best results for predicting the target variable value (i.e. crime rate). As a case in point, one key correlation we are working to build a predictive model for is the relationship between the geospatial region (i.e. district) and government spending in that region, and how the crime statistics (i.e. the target variable) in that region change. However, government spending is reported as the date the funding was authorized to be spent and the duration over which that funding is expected to be spent. Since we are looking to assess the impact of government spending on crime, determining to what degree a funding injection into a region is a complex problem in and of itself.

For example, does a funding of $1M dollars in that region impact a crime static 1 year (or one month, or one day) after the funding was executed? Our team will perform exploratory data analysis to make some assumptions about the temporal impacts of the data and what assumptions can be made, and what potential data reduction steps can be performed, given the association.

In the final data mining analysis step, our team will develop predictive models for crime statistics on a training set of data and evaluate the model performance on a reserved test data set. Our current plan is to develop a decision tree-based model, but we are open to explore other predictive machine learning models such as a Neural Network if time allows. We would also like to evaluate the performance of our models against other predictive crime metrics for NYC published for the same evaluation years.

V. ANTICIPATED OUTCOME

Our goal is to develop a model to accurately predict the expected crime rate for a specific geographic region given controllable or knowable input features such as total government spending in the region, spending, and the type of projects funded.

Intuitively, given that the objective of government spending in a given area is generally, but not always, to improve the quality of life of residents of the region, one would expect that there is an inverse correlation between spending and crime rate. That is, with increased funding an area will see decreased crime. Examples where this would not be the case would be the construction of a casino or the clearing of green areas (e.g., public parks) to facilitate the construction of a highway to improve traffic flow to another region. However, the data set we were able to obtain is specifically focused on funding for the revitalization of schools, which should have a direct positive impact on quality of life in a region.

Should our models prove successful in their ability to predict crime based on various funding patterns, it will be particularly interesting to explore whether the models are indicative of some previously unexplored ruleset that can inform local government officials how much funding is needed to curb crime and by how much they can expect crime to fall given a specific spending amount, focus area, and spending rate. Further, do the models provide an indication of diminishing returns? That is, is there a point at which the model indicates further spending will not further curb crime?

Our team expects to explore possible answers to these questions.

REFERENCES

[1] Hoffman, I., Evan, M. Heterogeneity in the Effect of Federal Spending on Local Crime: Evidence from Causal Forests. (2019, January). Cornerstone Research, Stanford University.

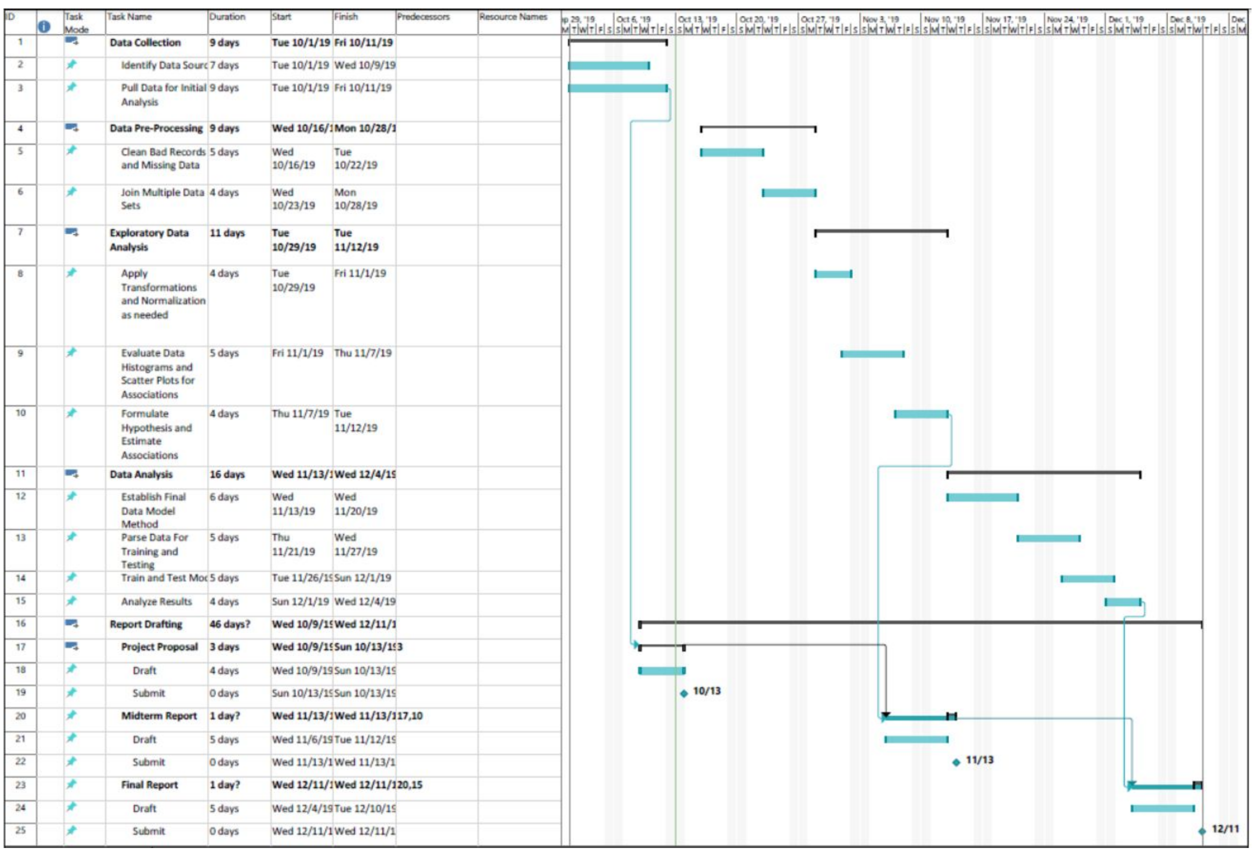
[2] Kneebone, E., Raphael, S. (2011, May). City and Suburban Crime Trends in Metropolitan America. Metropolitan Policy Program.

[3] Dunlap, E., Golub, A., Johnson B. (2006). Smoking Marijuana in Public: The Spatial and Policy Shift in New York City Arrests, 1992–2003.

[4] Grawert, A., Onyekwere, A., Kimble, C. (2018) Crime and Murder in 2018: A Preliminary Analysis. Brennan Center for Justice.

[5] FBI UCR, 2018 Crime in the United States https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018)

APPENDIX A



ID	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
1		**Data Collection**	**9 days**	**Tue 10/1/19**	**Fri 10/11/19**		
2		Identify Data Sourc	7 days	Tue 10/1/19	Wed 10/9/19		
3		Pull Data for Initial Analysis	9 days	Tue 10/1/19	Fri 10/11/19		
4		**Data Pre-Processing**	**9 days**	**Wed 10/16/1**	**Mon 10/28/1**		
5		Clean Bad Records and Missing Data	5 days	Wed 10/16/19	Tue 10/22/19		
6		Join Multiple Data Sets	4 days	Wed 10/23/19	Mon 10/28/19		
7		**Exploratory Data Analysis**	**11 days**	**Tue 10/29/19**	**Tue 11/12/19**		
8		Apply Transformations and Normalization as needed	4 days	Tue 10/29/19	Fri 11/1/19		
9		Evaluate Data Histograms and Scatter Plots for Associations	5 days	Fri 11/1/19	Thu 11/7/19		
10		Formulate Hypothesis and Estimate Associations	4 days	Thu 11/7/19	Tue 11/12/19		
11		**Data Analysis**	**16 days**	**Wed 11/13/1**	**Wed 12/4/1**		
12		Establish Final Data Model Method	6 days	Wed 11/13/19	Wed 11/20/19		
13		Parse Data For Training and Testing	5 days	Thu 11/21/19	Wed 11/27/19		
14		Train and Test Moc	5 days	Tue 11/26/1	Sun 12/1/19		
15		Analyze Results	4 days	Sun 12/1/19	Wed 12/4/19		
16		**Report Drafting**	**46 days?**	**Wed 10/9/1**	**Wed 12/11/1**		
17		**Project Proposal**	**3 days**	**Wed 10/9/1**	**Sun 10/13/1**	**3**	
18		Draft	4 days	Wed 10/9/19	Sun 10/13/1		
19		Submit	0 days	Sun 10/13/1	Sun 10/13/1		
20		**Midterm Report**	**1 day?**	**Wed 11/13/1**	**Wed 11/13/1**	**17,10**	
21		Draft	5 days	Wed 11/6/19	Tue 11/12/1		
22		Submit	0 days	Wed 11/13/1	Wed 11/13/1		
23		**Final Report**	**1 day?**	**Wed 12/11/1**	**Wed 12/11/1**	**20,15**	
24		Draft	5 days	Wed 12/4/19	Tue 12/10/1		
25		Submit	0 days	Wed 12/11/1	Wed 12/11/1		

APPENDIX B

Table 1: An Approximate Mapping of NYPD Precincts to NYC Community Districts

NYPD Prec.	Comm.District	Neighborhoods
		Manhattan
1	1	Battery Park, Tribeca
5	3	Lower East Side, Chinatown
6	2	Greenwich Village, Soho
7	3	Lower East Side, Chinatown
9	3	Lower East Side, Chinatown
10	4	Chelsea, Clinton
13	6	Stuyvesant Town, Turtle Bay
14	5	Midtown Business District
17	6	Stuyvesant Town, Turtle Bay
18	4	Chelsea, Clinton
19	8	Upper East Side
20	7	West Side, Upper West Side
22		Central Park
23	11	East Harlem
24	7	West Side, Upper West Side
25	11	East Harlem
26	9	Manhattanville, Hamilton Heights
28	10	Central Harlem
30	9	Manhattanville, Hamilton Heights
32	10	Central Harlem
33	12	Washington Heights, Inwood
34	12	Washington Heights, Inwood
		The Bronx
40	1	Melrose, Mott Haven, Port Morris
41	2	Hunts Point, Longwood
42	3	Morrisania, Crotona Park East
43	9	Soundview, Parkchester
44	4	Highbridge, Concourse Village
45	10	Throgs Neck, Co-op City, Pelham Bay
46	5	University Heights, Fordham, Mt. Hope
47	12	Wakefield, Williamsbridge
48	6	East Tremont, Belmont
49	11	Pelham Parkway, Morris Park, Laconia
50	8	Riverdale, Kingsbridge, Marble Hill
52	7	Bedford Park, Norwood, Fordham
		Brooklyn
60	13	Coney Island, Brighton Beach
61	15	Sheepshead Bay, Gerritsen Beach
62	11	Bensonhurst, Bath Beach
63	18	Canarsie, Flatlands
66	12	Borough Park, Ocean Parkway
67	17	East Flatbush, Rugby, Farragut
68	10	Bay Ridge, Dyker Heights
69	18	Canarsie, Flatlands
70	14	Flatbush, Midwood
71	9	Crown Heights South, Wingate
72	7	Sunset Park, Windsor Terrace
73	16	Brownsville, Ocean Hill
75	5	East New York, Starrett City
76	6	Park Slope, Carroll Gardens
77	8	Crown Heights North
78	6	Park Slope, Carroll Gardens
79	3	Bedford Stuyvesant
81	3	Bedford Stuyvesant
83	4	Bushwick
84	2	Brooklyn Heights, Fort Greene
88	2	Brooklyn Heights, Fort Greene
90	1	Williamsburg, Greenpoint
94	1	Williamsburg, Greenpoint
		Queens
100	14	The Rockaways, Broad Channel
101	14	The Rockaways, Broad Channel
102	9	Woodhaven, Richmond Hill
103	12	Jamaica, St. Albans, Hollis
104	5	Ridgewood, Glendale, Maspeth
105	13	Queens Village, Rosedale
106	10	Ozone Park, Howard Beach
107	8	Fresh Meadows, Briarwood
108	2	Sunnyside, Woodside
109	7	Flushing, Bay Terrace
110	4	Elmhurst, South Corona
111	11	Bayside, Douglaston, Little Neck
112	6	Forest Hills, Rego Park
113	12	Jamaica, St. Albans, Hollis
114	1	Astoria, Long Island City
115	3	Jackson heights, North Corona
		Staten Island
120	1	Stapleton, Port Richmond
122	2	New Springville, South Beach
123	3	Tottenville, Woodrow, Great Kills
		Other Precincts
		NYC Transit
		NYC Housing
		Park Police
		NYPD Headquarters
		Triboro Bridge/Tunnel
		Port Authority of NY/NJ
		Metro Transit