Hank Behaeghel
July 10th, 2023
AS.110.125

An Examination of NYC Shootings and Potential Budget Implications

Abstract

The following case studying revolves around New York City. The primary focus of this case study was New York City shooting data and whether murders resulted in shootings. A cursory exploration of data slowly developed into seeing if there was any correlation between demographic characteristics and the result of a shooting. The various variables of interest included a mix of demographic, location, and funding data. A logistic regression was employed as the model of choice due to the binary nature of the outcome either being a murder or not. The results were mixed in terms of providing any sort of predictive nature outside of age. Historical funding data was employed in a second reduced regression which, yielded trivial results.

Introduction

Given the rise of gun violence in America, we were interested in looking at New York City given that it is one of the more famous cities in the world and the city provides a wealth of historical data on shootings. Kindled by this rise of gun violence on a mass scale, gun legislation has come to the forefront of politics and policy decisions. This motivated us to try and identify potential demographic and geographic predictors for shooting outcomes. Due to limitations of the data set employed by this case study, and policy recommendations made in this study will not include a discussion on gun magazine capacity, which has been shown to examine in other literature.[1]

Murder is one of the most expensive crimes in terms of cost to perpetrators and victims.[2] This makes murder an interesting outcome to study due to its measurable detriment to society. Measuring the cost of crime is difficult and has been covered in numerous studies; varying approaches have yielded different results; however, a comprehensive study in 2010 demonstrated the social cost of murder to be $8.9million[3]. This drove the decision include a discussion of public finance within this case study.

Our examination of public finance was not as straightforward as we had hoped. The first avenue of public finance analysis relied heavily on payroll data that was extensive. However, despite being extensive it had its drawbacks as it did not include capital outlays. This led to a search for a comprehensive data set that included capital outlays for agencies in New York City as well as prior year adjustments. We were able to find such a data set that provided more reliable data for our second regression descried later. Regrettably, it meant that our efforts on the payroll data yielded little in the way of meaningful analysis beyond speculative proportional estimates for respective borough funding.

---

[1] Koper, et. al. 2009
[2] McCollister et. al. 2010
[3] Ibid.

Data

The data that was used for this case study was sourced from the New York City's public access data. Using NYC Open Data, we were able to find historical data on New York City's annual payroll data as well as shooting incidents reported to New York Police Department. New York City Independent Budget Office's site furnished us with historical agency expenditures from 1980 to 2022.

The historical payroll data spanned from 2014 to 2022. It included 5.11 million observations on employee data. The data set was robust but, included information that would not be relevant to my investigation. The names of employees were disregarded due to concerns over privacy. The main variables of interest in this data set were fiscal year, gross pay, OT pay, other pay, base salary, agency name, work location borough. These variables allowed for a brief analysis on which sectors of public service received the most funding from the city. In order to better understand the funding a new variable was constructed from the data within the payroll data, named 'Pay', it summed both the gross pay and the OT Pay as from the fact sheet provided with the data the gross pay did not include overtime. As this data is public there are some concerns given that the employee names are included in the raw data. They were then removed during the data cleaning process to mitigate any risk of impropriety.

New York Police Department reports their shooting data in a continuously updated set. The raw data set that was used for this case study had been updated on April 27th of 2023. This would allow for analysis to include shootings until the 31st of December 2022. The main variables of interest in this data set were occurrence date, boro, precinct, statistical murder flag, victim age, victim race, victim sex. The statistical murder flag was represented by a Boolean true or false. True indicated that the shooting resulted in a murder, meaning that the victim died as a result; and false for having survived the shooting. The variables of interest were subset from the data to create a slimmer data set.

From both sets of data, the Fiscal Year 2016 was chosen as training set. The reason for this is that FY 2016 in New York runs from July 1st of 2015 to June 30th of 2016. During this time frame there were no major global pandemics such as Covid-19 and the election of 2016 had not occurred yet. The reason this is important is because it allows for a subset of data that was not subject to large exogenous shocks that might have skewed the data out of a normal range. In addition, there were no major national legislations passed regarding firearms enacted during this time frame.[4]

Methodology

Data Cleaning

The software used for the data analysis and modeling in this case study was R Studio version 2023.06.01 *Mountain Hydrangea* running R 4.3.1 *Beagle Scouts*. R was chosen for its ability to handle larger data sets than Excel. Usually this would not be a problem however the payroll data

[4] H.R.3411 was passed regarding gun legislation but, only affected reporting between agencies regarding background checks.

contained over 5.1 million rows (observations) and Excel is limited to just over a million rows per worksheet. The shooting data would have been easily handled by Excel however, to centralize all work and analysis both sets were examined in R.

The main packages used for this analysis were “Tidyverse” and “Conflicted”. The “Tidyverse” package is well known in data analytics as it is intuitively programmed such that functions often match the verb that one might use to describe the operation.[5] The reason for using the “Conflicted” package was to allow for an easy reconciliation between the ‘filter’ function that exists in both base R and “Tidyverse”, “Conflicted” allows one to preempt future occurrences of the error by indicating a usage preference. Since most of the analysis relies heavily on the “Tidyverse” package, its respective ‘filter’ function was indicated as preferred.

The main hurdle for the data cleaning was the removal of variables that were not pertinent to the analysis. The data was subset to include the variables of interest, that are both outlined in the introduction as well as the coding files. Here, another benefit of using R arose. In R you can maintain original datasets in the global environment and continually reference them in an easier manner than Excel since one can explicitly reference with just the name of the variable. As a result each iteration of sub setting the data created a new variable explicitly named. Any aesthetic changes to the final subsets were clobbered to the same variable name as is seen in the code.

The data set that required the most work was the shooting data. Due to factors not listed in the data set fact sheet but could reasonably be presumed to be due to human error some observations were missing values for our variables of interest. Due to the categorical nature of the variables chosen for the analysis observations that were missing values were removed entirely. In this data set often these values appeared as “UNKNOWN” or simply “U”. The reasoning behind the removal of all observations with unknown within the shooting data was that only 126 observations out of 27,312 were removed. This left adequate data to complete the case study. In addition to removing observations with unknowns the “Murder” variable was recoded as 1 for murder, 0 otherwise to replace the Boolean T/F in order easier model interpretation.

The payroll data required small changes to prove useful in analysis. Agencies were determined to be one of three categories: police, education, social services. The reason behind the categories is that the three are often talked about in unison in policy debates on how to reduce crime. Think of the defund the police discourse that has arisen. The rationale behind this is that the money should be used in other avenues to address the systematic root causes of crime.[6] Therefore all agencies were examined and nine chosen as encompassing. For education the Department of Education which included differing pay codes for: hourly support staff, paraprofessionals, pedagogical (“tenured” teachers), per diem teachers, and per session. For social services the following were included, Department of Parks and Rec, Department of Homeless Services, Department of Social Services. For police solely the police department was used; any Department of Justice related positions such as District Attorney were not included because their direct involvement in physical occurrences of crime is limited.

[5] Tidyverse.org
[6] Defundthepolice.org

Each data set was subset on dates from July 1st 2015 to June 30th of 2016 to match FY 2016 for New York City's municipal budget. This was done for that reason that these dates exclude any major exogenous shocks that were previously discussed in the *Data* section. From here the analysis ensued.

Models

The following model was the first model used. It was selected to be a logistic model for the reasoning that murder is a binary variable. Either a shooting resulted in a murder, or it did not. As a result, murder could be coded as 1 or a 0. This reveals a probabilistic outcome and therefor a logistic regression is warranted as opposed to a general linear model.[7] The result was that we could see the contribution of our variables of interest to the likelihood of getting shot. The first model used on the shooting data was as follows:[8]

$$\log\left[\frac{P(\text{Murder}=1)}{1-P(\text{Murder}=1)}\right] = \alpha + \beta_1(\text{BORO}_{\text{QUEENS}}) + \beta_2(\text{BORO}_{\text{BRONX}}) + \beta_3(\text{BORO}_{\text{BROOKLYN}}) + \beta_4(\text{BORO}_{\text{S.I.}}) + \beta_5(\text{VIC\_AGE\_GROUP}_{\text{45-64}}) + \beta_6(\text{VIC\_AGE\_GROUP}_{\text{18-24}}) + \beta_7(\text{VIC\_AGE\_GROUP}_{\text{25-44}}) + \beta_8(\text{VIC\_AGE\_GROUP}_{65+}) + \beta_9(\text{VIC\_SEX}_{\text{M}}) + \beta_{10}(\text{VIC\_RACE}_{\text{BLACK}}) + \beta_{11}(\text{VIC\_RACE}_{\text{WHITE HISPANIC}}) + \beta_{12}(\text{VIC\_RACE}_{\text{BLACK HISPANIC}}) + \beta_{13}(\text{VIC\_RACE}_{\text{ASIAN / PACIFIC ISLANDER}})$$

We then preformed a second logistic regression. The main difference between the two models is that the second model has reduced the levels of the first model and includes historical agency funding. Both were done using the same shooting dataset. The data set for the second model was a clone of the first. This cloned data set was then merged with the historical expenditure data based on fiscal year in which each shooting occurred.

$$\log\left[\frac{P(\text{Murder}=1)}{1-P(\text{Murder}=1)}\right] = \alpha + \beta_1(\text{VIC\_AGE\_GROUP}_1) + \beta_2(\text{VIC\_SEX}_{\text{M}}) + \beta_3(\text{VIC\_RACE}_1) + \beta_4(\text{EDU}) + \beta_5(\text{PD}) + \beta_6(\text{SOCIAL})$$

Variables

The variables used in this regression were: borough (coded as BORO), victim's age group, sex of the victim, and the victim's race. The first four coefficients represent the change in likelihood of a shooting resulting in a murder based on the borough it occurred. The next four represent the change in likelihood of a shooting resulting in a murder based on the victim's age. The data set that I used only reported this in tranches, so the exact age of victims is not reported. The coefficient β_9 represents the change in likelihood of a shooting resulting in a murder based on the victim's sex, specifically men in this case since men are regressed against female based on the levels specified to the R code. The remaining coefficients are displaying the change in likelihood of a shooting resulting in a murder based on the victim's race.

[7] Provost and Fawcett, Chapter 5

[8] I sincerely apologize for how messy it looks, even LaTeX could not save it.

Despite its rather brutish form, it provides useful information on potential demographic predictors of being murdered in a shooting. It is important to note at this point that the finer geographical data include in the raw data could be used to do an analysis regarding likely response time which could influence outcomes but, this is beyond the scope of this study.

In the second model a majority of the variables are the same. The main difference is that in the cloned data set the levels for each of the demographic variables were reduced to two levels. Now the age group represents whether an individual was an adult (1) or not (0), white (0) or a person of color (1). The variables, EDU, PD, and SOCIAL represent the expenditures for each of these departments each year and the coefficients represent the change in probability of a shooting resulting in a murder based on an increase of funding.

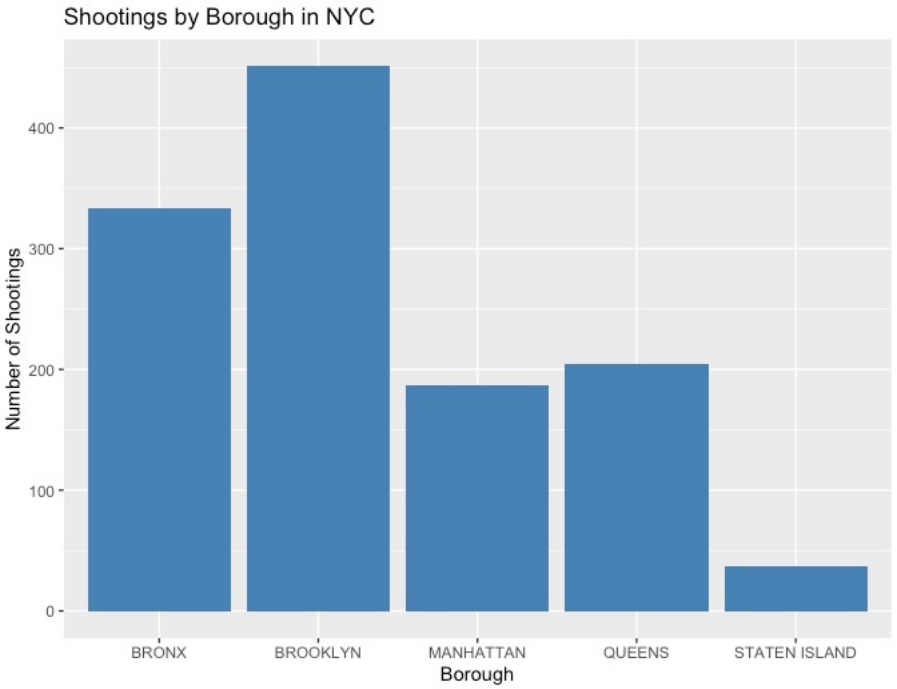
Results

The following table displays the results of the first logistic regression done on the demographic data included in the shooting data. The levels of significance are 0.001, 0.01, 0.5, represented by a decreasing number of '*'. The comparison case or compared to a dummy of '0' for the respective variables of borough, age, sex, race is: an individual form Manhattan, aged less than 18 years old, female, white. This was chosen since R takes factor variables and removes one level to regress to. I assumed that a white female child from Manhattan would likely not be exposed to many shooting opportunities and therefore as a comparison might reveal more correlation between being murdered and varying demographic indicators. There are several observations that can be made from these results. The first being that it appears that race has no significant correlation with the result of the shooting being a murder or not. The second is that age has the greatest correlation with death. This could be explained by the simple fact that as one gets older the body is less resilient. A notable result is that Brooklyn despite being the location of a disproportioned number of shootings shows no statistically significant correlation with the outcome of the shooting.

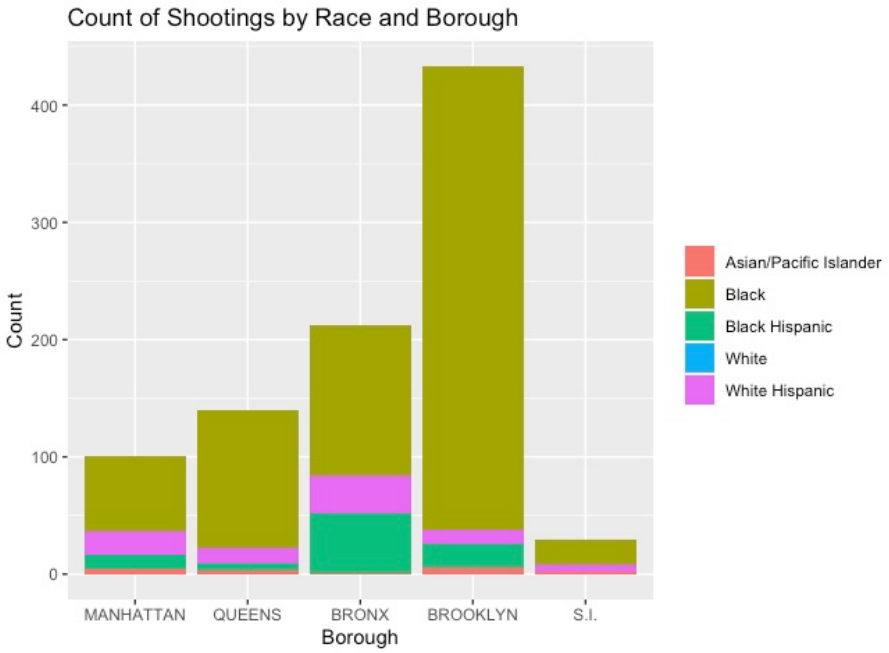
	Model 1
(Intercept)	-1.66 **
	(0.55)
BOROQUEENS	-0.34
	(0.30)
BOROBRONX	0.17
	(0.25)
BOROBROOKLYN	0.39
	(0.24)
BOROS.I.	-0.02
	(0.45)
VIC_AGE_GROUP45-64	0.83 *
	(0.37)
VIC_AGE_GROUP18-24	0.12
	(0.30)
VIC_AGE_GROUP25-44	0.60 *
	(0.29)
VIC_AGE_GROUP65+	2.37 *
	(0.93)
VIC_SEXM	-0.25
	(0.24)
VIC_RACEBLACK	-0.14
	(0.42)
VIC_RACEWHITE HISPANIC	0.09
	(0.44)
VIC_RACEBLACK HISPANIC	-0.36
	(0.49)
VIC_RACEASIAN / PACIFIC ISLANDER	0.50
	(0.70)
N	1260
AIC	1248.20
BIC	1320.15
Pseudo R2	0.04

*** p < 0.001; ** p < 0.01; * p < 0.05.

This graph shows that this is the opposite of one might expect, as Brooklyn had the highest number of shooting incidences. These would indicate that perhaps age is the largest contributor to a shooting resulting in a murder.[9] All of these age groups occur post state funded education. This could indicate that educational programs lack the ability to properly prepare individuals to avoid being victims of violent crime (this could of course be random).



I would have expected that race might have displayed more correlation based on the distribution of races in shooting incidents. Here we see that Black individuals appear to be victims in larger proportion than their White, Asian, or Hispanic counterparts. This is counter to the fact that the Black population of New York represents ~20% of the population.[10] One would expect that the victims of murders would mirror the population distribution if they were randomly distributed. It would seem that they are not after both a qualitative and quantitative analysis are performed. This would imply that there is a factor beyond the scope of this case study that leads to the skewed representation of minorities in shootings.



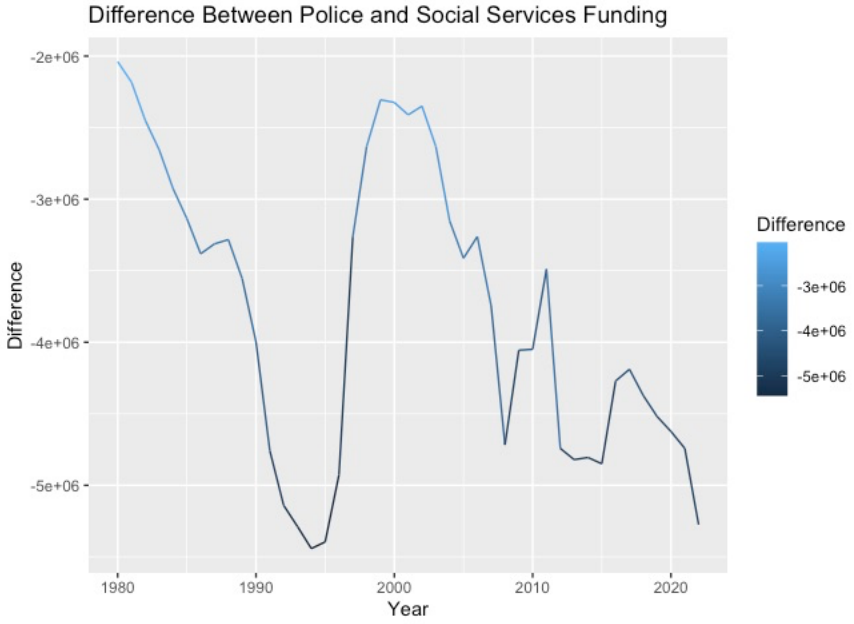
[9] I cannot stress enough that this does not imply causality. The case study as of now does nothing to establish such a relationship but, rather tries to provide a structured analysis to suggest better funding distribution.
[10] New York City Department of Planning

Further exploration was done using another logistic regression. This model was constructed with a merged data set between the shooting data and the historical funding data. The results are in the table seen here. Important differences between this model and the last one is that in the new data set the levels of factors were reduced; age and race were both reduced to two levels. The decision for this was motivated by the results of the last model. There was statistical significance for each of the age groups and therefore muddied any signal. The results for race were not statistically significant so, I choose to make the distinction between White and POC. The results were that being, and adult increases the probability of a shooting resulting in a murder and being a person of color reduces that probability. Disappointingly, there was not much in the way for effects due to funding. I suspect that this is because the shooting data only goes back to 2006 and department funding does not change shooting to shooting but, rather year over year.

	Model 1
(Intercept)	-0.60 **
	(0.21)
VIC_AGE_GROUP1	0.53 ***
	(0.06)
VIC_SEXM	-0.09
	(0.05)
VIC_RACE1	-0.43 ***
	(0.09)
EDU	0.00 ***
	(0.00)
PD	-0.00
	(0.00)
SOCIAL	-0.00 ***
	(0.00)
N	26343
AIC	25714.77
BIC	25772.02
Pseudo R2	0.01

*** p < 0.001; ** p < 0.01; * p < 0.05.

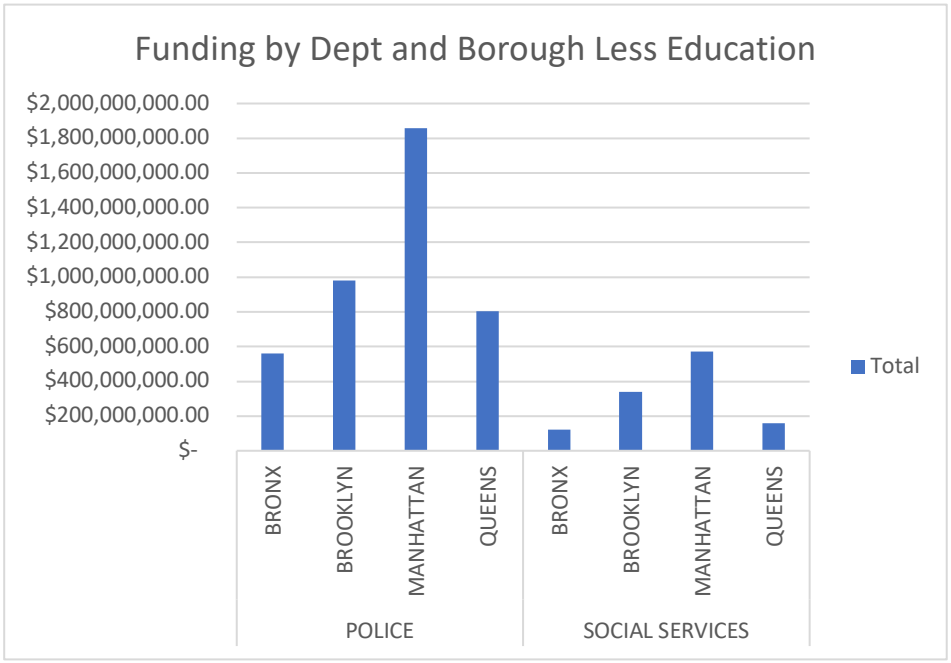
Analysis was done on the funding for the New York Police Department, Department of Social Services, and Department of Education. I was curious because of a paper written by Sara Heller and coauthors that examined youth programs that lead to lower incarceration rates and violent crime arrests.[11] I chose to examine the difference between police funding and social services fund; the reason for this is because New York City's public school system is more expansive than any other and therefore education spending far outpaces police and social services.[12] The following chart shows the difference. The negative values indicate that the Department of Social Services received more funding than NYPD.



[11] Heller et. al. 2017

[12] At one point 1 in 100 Americans was a NYPS student.

Using the payroll data we were able to do some qualitative analysis on which boroughs receive the most funding from the city for employees. What we found was that on payroll only out of the agencies we thought relevant to our analysis, Manhattan was the borough that receives the most money. This is interesting considering it is not the most populous of New York's five boroughs; Brooklyn is the most populous, with roughly ~1 million more inhabitants than Manhattan for the past thirty years.[13] Also, interesting to note is that on this graph it would appear that the police receive more funding than social services which runs contrary to the graph prior on the difference in funding between the two departments. This is likely due to the fact that the payroll data is more granular than the historical data and when selecting agencies of interest some did not make the cut. However, this graph remains useful in understanding that despite differing from the historical results we can see that the distribution of funds across boroughs is similar and that it does not reflect population. This implies that there may likely be political agendas skewing funding or perfectly innocuous reasons not explored in this case study.



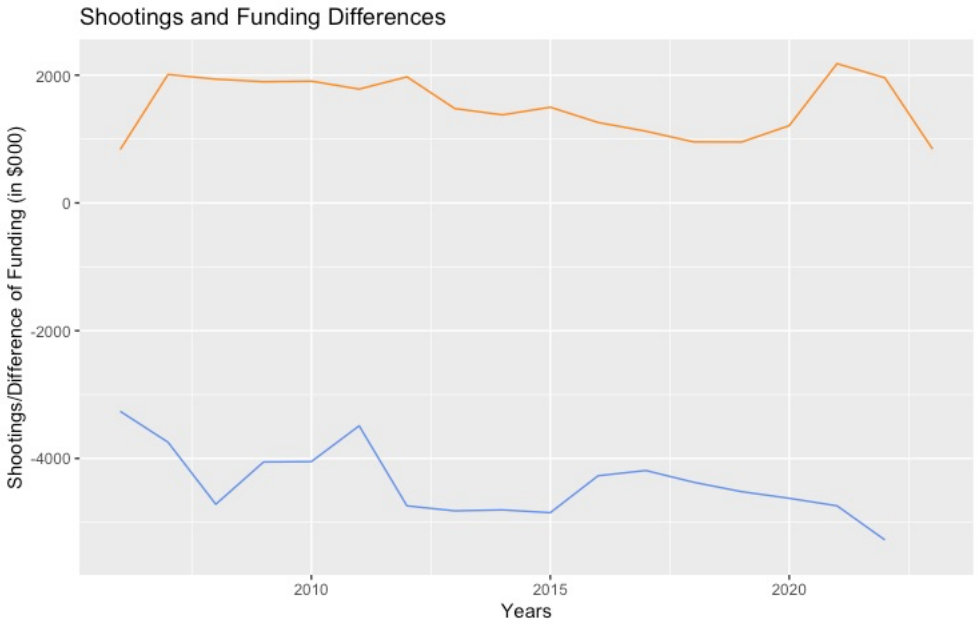
Conclusion

Commentary

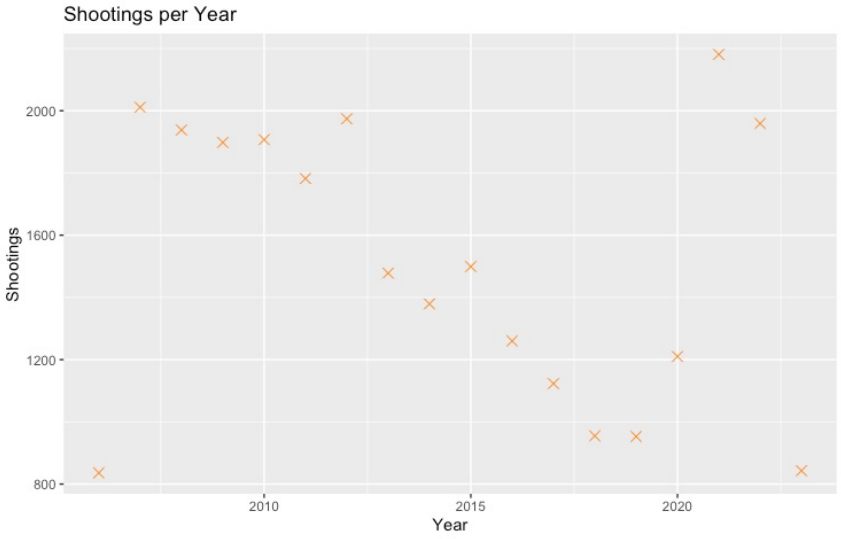
The results of this case study were quite interesting. As we saw in the examination of funding, despite the popular discourse it would appear that the police is funded less than social services and that neither had any correlation with the outcomes of shootings. This was contrary to our expected results. A large part of this study was motivated by the results of Heller and her coauthors paper on the investment in after school programs and youth involvement programs. We would have expected these potential effects to be captured in the coefficients on our education and social services funding variable. There were limitations such that the amount of funding only has sixteen variations because the shooting data only went as far back as 2006. Despite this we were still able to find encouraging results that despite inspiration for this topic

[13] US Census Bureau, compiled by citypopluation.de

being the rise of mass shootings in America, it would appear that at least in New York children for the most part are not victims.[14]



An additional point of interest is that if one compares the difference in police and social services funding with the number of shootings per year it will appear as if they have a roughly parallel trend at times. This implies that increasing the funding of social services compared to police services does reduce shootings.[15] Which as previously outlined would greatly be of benefit. Murder has a cost of roughly $8.9million and if small increments of funding to social services relative to police funding decreases the number of shootings its best to do so from a policy perspective. If the shooting data extended further, it would be more conclusive. This graph displays this thought, here the blue line is the same from the prior graph on police spending; the orange line represents the number of shootings per year.[16] This only modification was to divide the difference by a thousand to better see trends and allows for the magnitudes to be comparable.



Another interesting conclusion from this graph bis that the past handful of years have seen a spike of shootings. Statistically they may not be considered outliers, however they are outliers of the trend of decreased shootings over time. This interesting because the spike is during the years of pandemic restrictions. This could be an area of further study. The absence of gross outliers does assuage any fear that any one year is swaying the data one way or the other.

[14] This comes from the interpretation of a significant and positive coefficient in the VIC_AGE_GROUP variable in the reduced regression.

[15] Again, not enough to establish causality but, food for thought.

[16] I fought with R for so ling yet no legend would generate, I apologize for this.

Future Considerations

Our original goal was to see whether we could find significant correlations between demographic, regional, and funding data on the outcomes of shootings. What we found was that on a granular level there is not much in the way of statistical significance to be found. However, if one takes a step back, we see that adults and persons of color are more likely to die in a shooting incident. Our case study does not seek to offer any sort of causality as to why this is. This is the biggest question and a massive undertaking. As for this project itself, the shooting data included the GPS locations of the nearest street corner for each shooting. A next step had it been feasible in time would be to overlay these positions on a map of NYC school districts, parks, police precinct, and hospitals. One could potentially then to a nearest neighbor analysis or find some geographical factor beyond simply the borough of the shooting that may sway outcomes. We wish we could have done this in time but, current such an analysis is beyond the current programming skills.

One last consideration would be data. There was a great amount of reworking due to historical data that made the payroll data less usable and impactful in the case study. Despite being more accurate the historical agency expenditure data did not break down geographically which was a big loss to potentially exploitable variation. There might be a way to construct this perfect data set and to continue our analysis.

We would also like to consider that in Appendix A there are the results of a third regression that, adds location back into the reduced model. The results do show significance in all boroughs except for Staten Island, this might be due to limited shootings being reported here. This again does not offer much of a clear signal because it seems that the shootings are proportional to population of each borough and if all are significant it does not mean all that much. Nonetheless it may be a launching point for further study.

Works Cited

Agency Expenditures. IBO. (n.d.). https://ibo.nyc.ny.us/fiscalhistory.html

America. New York City Boroughs (USA): Boroughs - Population Statistics, Charts and Map. (n.d.). https://www.citypopulation.de/en/usa/newyorkcity/

Defund the police. Defund The Police. (2021, March 18). https://defundthepolice.org/

Heller, Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2017). Thinking, fast and slow?: Some field experiments to reduce crime and dropout in Chicago. The Quarterly Journal of Economics, 132(1), 1–54. https://doi.org/10.1093/qje/qjw033

H.R.3411 - 114th Congress (2015-2016): Fix Gun Checks Act of 2015. (2015, September 28). https://www.congress.gov/bill/114th-congress/house-bill/3411

Koper, C. S., Johnson, W. D., Stesin, K., & Egge, J. (2019). Gunshot victimisations resulting from high-volume gunfire incidents in minneapolis: Findings and policy implications. *Injury Prevention, 25*, i9-i11. doi:https://doi.org/10.1136/injuryprev-2017-042635

McCollister, K. E., French, M. T., & Fang, H. (2010). The cost of crime to society: new crime-specific estimates for policy and program evaluation. *Drug and alcohol dependence*, *108*(1-2), 98–109. https://doi.org/10.1016/j.drugalcdep.2009.12.002

New York City Department of City Planning | Population Division. (2023, June 23). *Dynamics of racial/Hispanic composition in NYC neighborhoods*. ArcGIS StoryMaps. https://storymaps.arcgis.com/stories/46a91a58447d4024afd00771eec1dd23

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data minig and data-analytic thinking*. O'Reilly.

Tidyverse packages. Tidyverse. (n.d.). https://www.tidyverse.org/packages/

Appendix A

Results from adding the BORO variable back into the second logistic equation. Noisy signal therefore was not included in the main analysis.

	Model 1
(Intercept)	-0.71 ***
	(0.21)
BOROQUEENS	0.13 *
	(0.06)
BOROBRONX	0.13 *
	(0.05)
BOROBROOKLYN	0.12 *
	(0.05)
BOROS.I.	0.18
	(0.10)
VIC_AGE_GROUP1	0.53 ***
	(0.06)
VIC_SEXM	-0.09
	(0.05)
VIC_RACE1	-0.42 ***
	(0.09)
EDU	0.00 ***
	(0.00)
PD	-0.00
	(0.00)
SOCIAL	-0.00 ***
	(0.00)
N	26343
AIC	25715.22
BIC	25805.18
Pseudo R2	0.01

*** p < 0.001; ** p < 0.01; * p < 0.05.