How Does Streetlighting Impact Night Crime?

Hasti Zahed
UNIVERSITY OF TORONTO | ECO225 | APRIL 2024

Introduction

This paper aims to explore the influence of streetlighting on night crime patterns in Boston zip codes. The presence of lighting may be an important deterrent for crime, as criminals generally operate with the intention of avoiding detection and apprehension. Low visibility environments may thus be associated with higher crime rates. While well-lit environments can provide low visibility for crimes like petty theft through crowd coverage, their effect is difficult to quantify. Hence, this paper defines visibility quite literally as darkness, looking at the relationship between artificial environmental lighting and night crimes. Focus is placed on artificial lighting in the form of streetlights however natural lighting— affecting hours of sunshine in a month and seasonal factors like precipitation— shall also be considered as they alter factors beyond human control (weather conditions, social schedules etc.). Including this analysis will help guide intuition when it comes to identifying potential confounding variables for regression analysis.

Previous literature into the impact of streetlights on crime patterns both in an observational and experimental manner has found varying results. Xu et al. (2018) find an inverse relationship between the presence of streetlights and the occurrence of crime across census block groups in Detroit. Chalfin et al. (2022) find a similar result through a randomised experiment assigning streetlights to different localities in New York. Temporal analyses through the consolidation of data from multiple different papers showed that more streetlights significantly reduced crime, but that this effect was larger for the UK than for the US (Welsh et al., 2008).

By contrast, a paper by Atkins et al. (1991) found no evidence to support an association between increased streetlighting and reduced crime. Similarly, Ken (1999) found that targeted increases in streetlighting had preventative effects but that general increases didn't have universal results.

Regarding natural lighting, a paper by Dominguez et. al (2023) found that later sunsets cause a reduction of criminal activity around dusk but crime patterns for the rest of the day stay relatively similar (before and after Daylight Savings Time).

These varying findings make this project worthwhile because the conclusion could go in either direction. The central goal of this research is to isolate potential effects of streetlighting from confounding variables, specifically income and population density in a zip code. Unlike previous literature, the isolation of night crimes allows for a spotlight to be placed on the impact of artificial light. While the sun and cyclical seasonal changes are beyond human control, urban planning and government funding into streetlight allocations are within the bounds of our capabilities. This research can provide a basis for the usefulness of streetlights as a night crime deterrent.

By consolidating streetlight locations with those at which crimes were reported, this project will explore the relationship between artificial lighting and crime patterns at the zip code level. An inverse relationship is hypothesised. Other fluctuating factors like income and educational differences between zip codes will be factored in to allow for further insight into whether lighting is the driver of crime trends.

The research has found that natural lighting conditions do not disproportionately affect crime patterns. Crime trends are almost equally distributed (with a slight favour to lighter hours) during the lighter and darker hours of the day, a trend that is consistent between seasons and years. Most criminal activity tends to occur in the Summer and Autumn. There is, however, a seasonal discrepancy between night crimes and day crimes. The former tend to happen in darker seasons (winter and autumn), while the latter occur more frequently in spring and summer. Analyses of income and streetlight locations showed that some of the zip codes with the highest income had greater streetlight presence compared to lower income zip codes. The majority of night crime occur in low-income districts with mediocre or weak streetlight density. This points towards wealth playing a role in how lighting conditions are related to criminal activity. Further analysis will attempt to control for income and other potentially different factors between districts to observe whether the effect of artificial lights among otherwise comparable districts is notable.

Data

The primary dataset is provided by the Boston Police Department (BPD) and includes information including but not limited to the type of crime, when it occurred, and its location (latitude, longitude) (Boston Police Department, 2019). Streetlight data is taken from a dataset provided by the City of Boston government website (Analyze Boston, 2016). It displays the locations (latitude, longitude) of all streetlights in the city. All additional variables (not originally in the dataset) were obtained from the US Census Bureau. This includes the TIGER shapefiles that were used to overlay latitudes and longitudes over Boston zip codes. The dependent variable is number of crime incidents per 1000 people in each zip code for every year and month for which data was available. The independent variables are presence of streetlighting, season, income, population, hours of sunshine, precipitation, and various demographic controls: educational attainment, race, age, and sex.

Number of crime incidents per 1k people in different zip codes and at different temporal intervals will be measured to observe crime patterns in the context of the varying presence of light and environmental factors. Analysis will be done both over the entire data and with focus on night crimes, defined by comparing the time that the crime occurred with average sunset/sunrise times in different months. These times were scraped for 2016 from the U.S Climate Data website. A single year was chosen to extrapolate from because average sunset/sunrise times did not change notably for corresponding months between years. 2016 was chosen specifically because it was a year with data for every month in the original dataset.

The independent variables each influence the street or criminal environment in some way. Variance in these variables will discern the potential existence of a relationship with the dependent variable.

Summary Statistics

Table 1.1: Summary Statistics of Night Crimes per 1k People in a Zip Code/Month/Year

Count	1211
Mean	4.706
Std	4.069
Min	0.0185
25%	2.123
50%	3.602

75%	6.004
Max	36.658

Table 1.2: Summary Statistics of Night Crimes in Each Year

Year	2015	2016	2017	2018
Count	18184	30788	30986	20904
Unique	22	24	25	22
Top	Theft	Theft	Theft	Theft
Freq	4053	6004	5723	3777

Table 1.1 shows that there exists a lot of variation in how frequently Boston experiences crime as the mean and standard deviation are almost identical. Different zip codes likely have different populations; those with more people may have higher occurrences of crime compared to sparsely populated areas. Table 1.2 displays summary statistics for night crimes in Boston in each year. 2015 starts from June and 2018 is missing data for November and December, hence why the counts are smaller compared to 2016 and 2017. This is something to be mindful of when interpreting year-specific results or making comparisons across years. Summary statistics over the entire dataset (including daytime) affirmed Theft as the most commonly occurring crime. Theft is defined as: auto theft, larceny, residential burglary, larceny from a motor vehicle, robbery, commercial burglary, and other burglary— defined by the BPD dataset and not elaborated upon.

Table 1.3: Summary Statistics of Boston Streetlights

count	73036
unique	47
top	02128
freq	4641

According to the streetlight dataset last updated in 2016, there are 73,036 streetlights in the city of Boston. Most of them are in the 02128 zip code. These summary statistics imply that the distribution of lighting across the city is unequal, with some zip codes having as few as one streetlight. This information will later be used to determine streetlight densities for different zip codes in order to see the distribution of artificial light across localities.

Table 1.4: Summary Statistics of Income

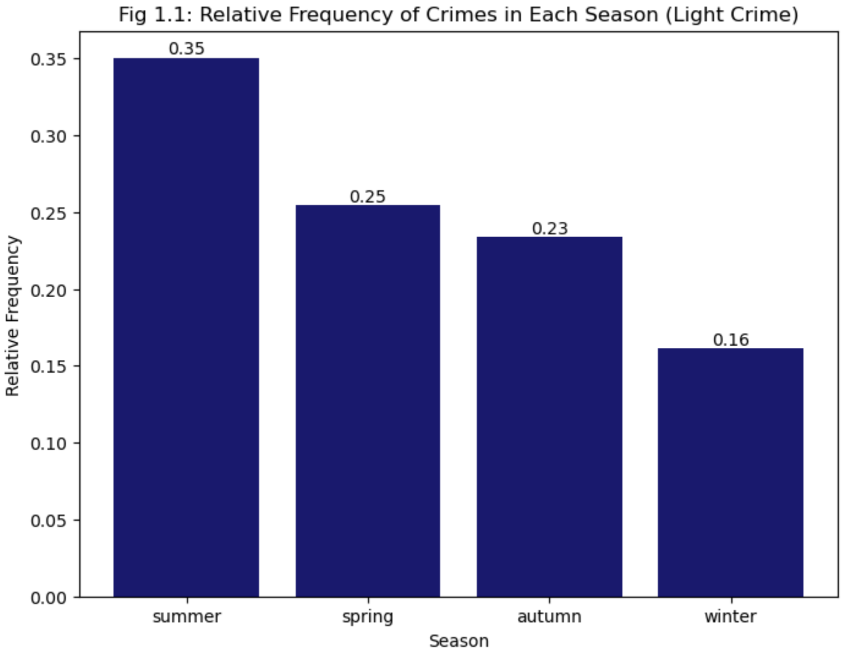
count	1211
mean	38908.40
std	21194.13
min	7070.00
25%	24119.00
50%	34419.00
75%	55417.00
max	92220.00

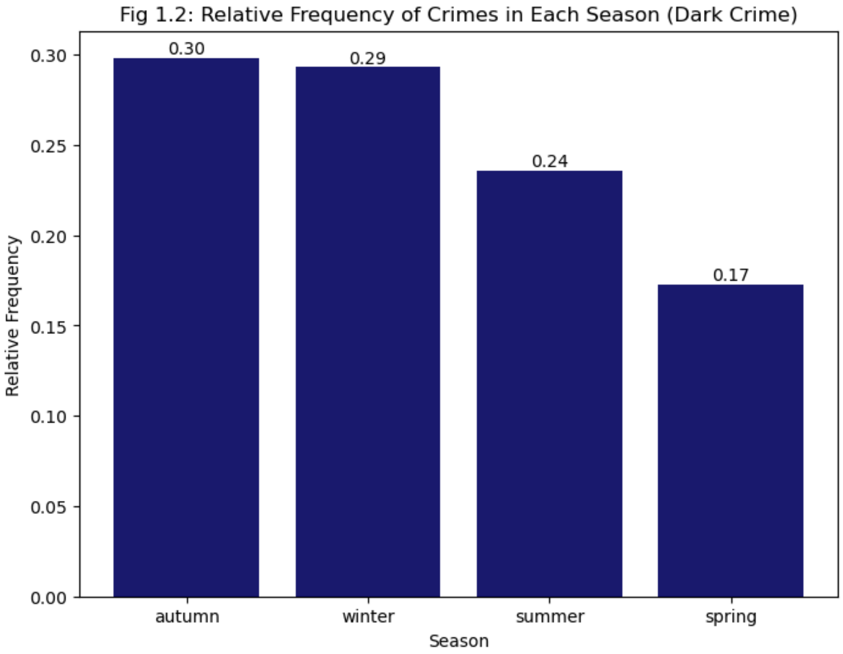
The high variation in income shown in Table 1.4 indicates that wealth is disproportionately concentrated. Since the mean is greater than the median, a positive skew is implied. This means that much of the wealth is Boston is concentrated in higher percentiles. Income inequality can contribute to criminal activity, but also to area funding which may explain greater or fewer streetlights in a zip code. Additionally, richer areas may experience and report more theft; the most common crime occurring in the dataset over all years.

Table 1.5: Summary Statistics of Seasons

Season	Autumn	Spring	Summer	Winter
Count	69676	60568	83149	56911
Unique	27	24	25	27
Top	Theft	Theft	Theft	Theft
Freq	13988	10691	16767	10989

These summary statistics communicate that in Boston from 2015-2018, most crimes occur in summer while fewest occur in winter. This result is interesting as it disputes the hypothesis that darkness and crime are positively related. However, it is important to note that environmental factors other than light may be at play here, since winter also brings harsh and uncomfortable weather. Cold weather may, itself, act as a deterrent for crime since it adds a layer of discomfort to illegal activity.





The relationship between crime and seasons is meaningful when inquiring about the influence of natural lighting on crime. Summer is generally a lighter season, with longer days and more days with sunlight. In the winter, days are shorter, and the sun is seen relatively less. The Boston data from 2015-2018 observes interesting differences in the temporal trends of night and day crime. While day crime seems to occur most frequently in Summer and Spring, night crime occurs more frequently in Winter and Autumn. An intuitive explanation for this is that daytime in summer and spring observes lively streets, the ideal environment for petty crimes like theft. Low visibility can immediately be thought of as synonymous with darkness however crowded areas may also provide cover for illegal activity. Additionally, summer vacation may see an increase in the number of vacant homes that can be targeted in broad daylight. Winter and autumn, however, see fewer people on the streets during the day and popular schooling schedules prevent mass vacationing. Hence, night crime may be higher because criminals rely more on low visibility from darkness when there isn't a crowd to get lost in.

In the context of lighting, this both supports and refutes the idea that more natural lighting is congruent with fewer crimes. However, in an environmental and social context these results are unsurprising. Winter entails snow and a higher likelihood of blocked, icy, or wet roads. This could contribute to more vehicle accidents, especially in low visibility. The social aspect has been discussed above.

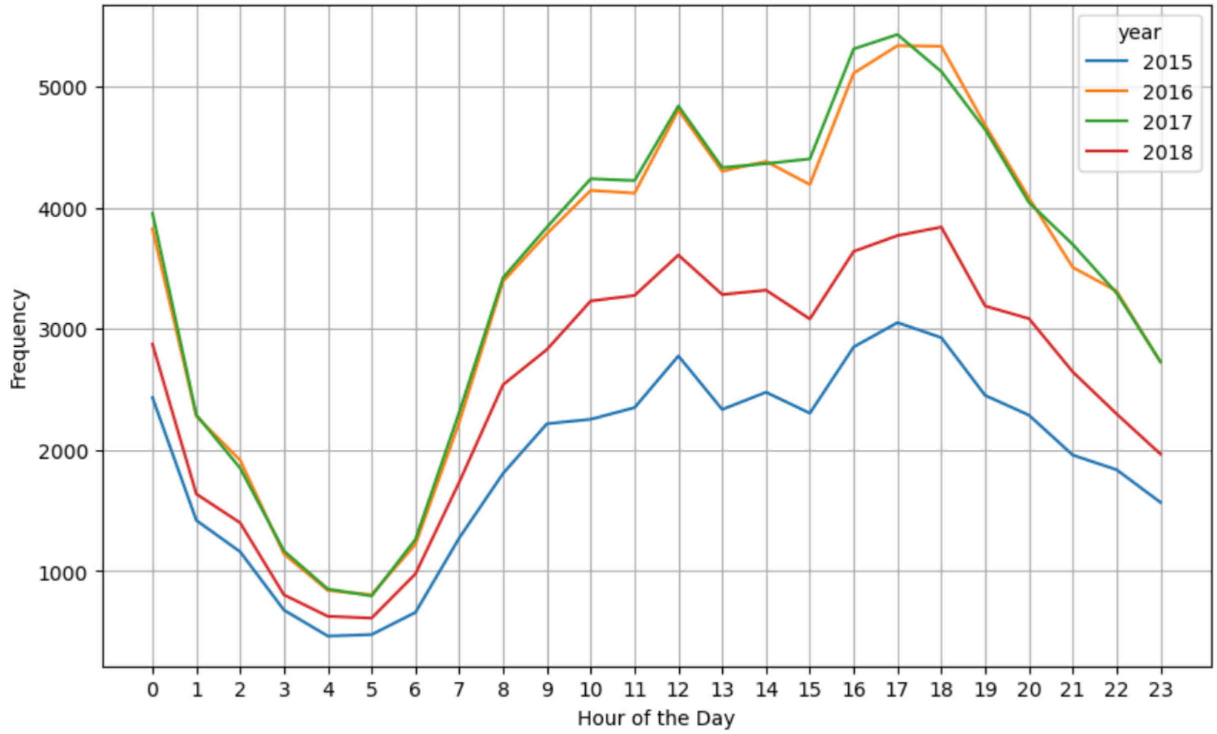


Fig 1.3: Frequency of Crime Occurences by Time of Day in Boston from 2015-2018

Figure 1.3 disputes the hypothesis that lower visibility is associated with higher crime, but it is important to remember that these are natural lighting conditions. While more crime does seem to occur in lighter hours— a result that is consistent across seasons, it may be a result of higher social activity during the day. Light and dark hours are defined by comparing the time of occurrence with the average sunrise and sunset times in a particular month.

Criminals are human so it can be assumed that they mostly follow a common circadian rhythm. Comparing the findings of criminal patterns under natural light to those of artificial light may be able to isolate the effect of lighting, unaffected by human social patterns. Furthermore, the dataset is taken from the Boston Police Department. It may be possible that police are more efficient at catching and documenting criminal activity under high visibility conditions. Alternatively, people may only realise their victimhood the day after the crime has occurred (e.g. vehicle theft). Day crime may thus be overrepresented in the data.

Fig 1.4: Shooting Incidents for Crimes in Boston (2015-2018)

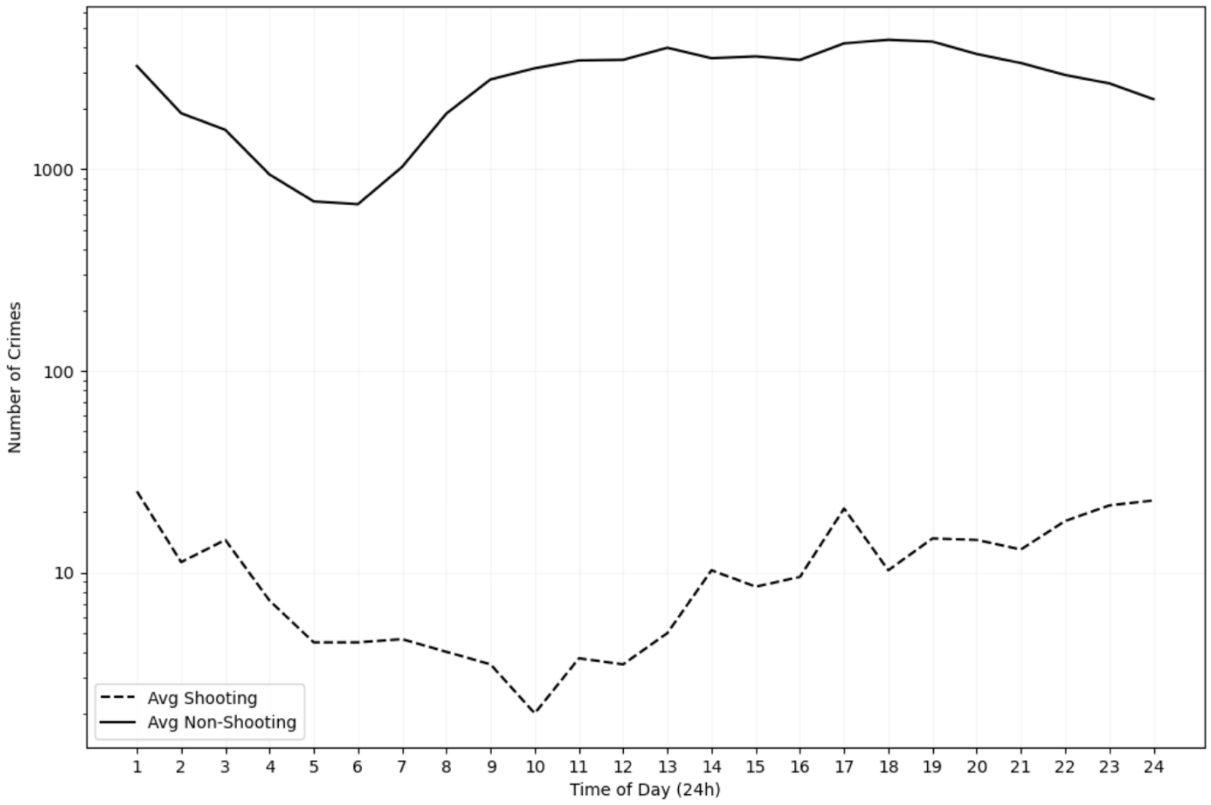
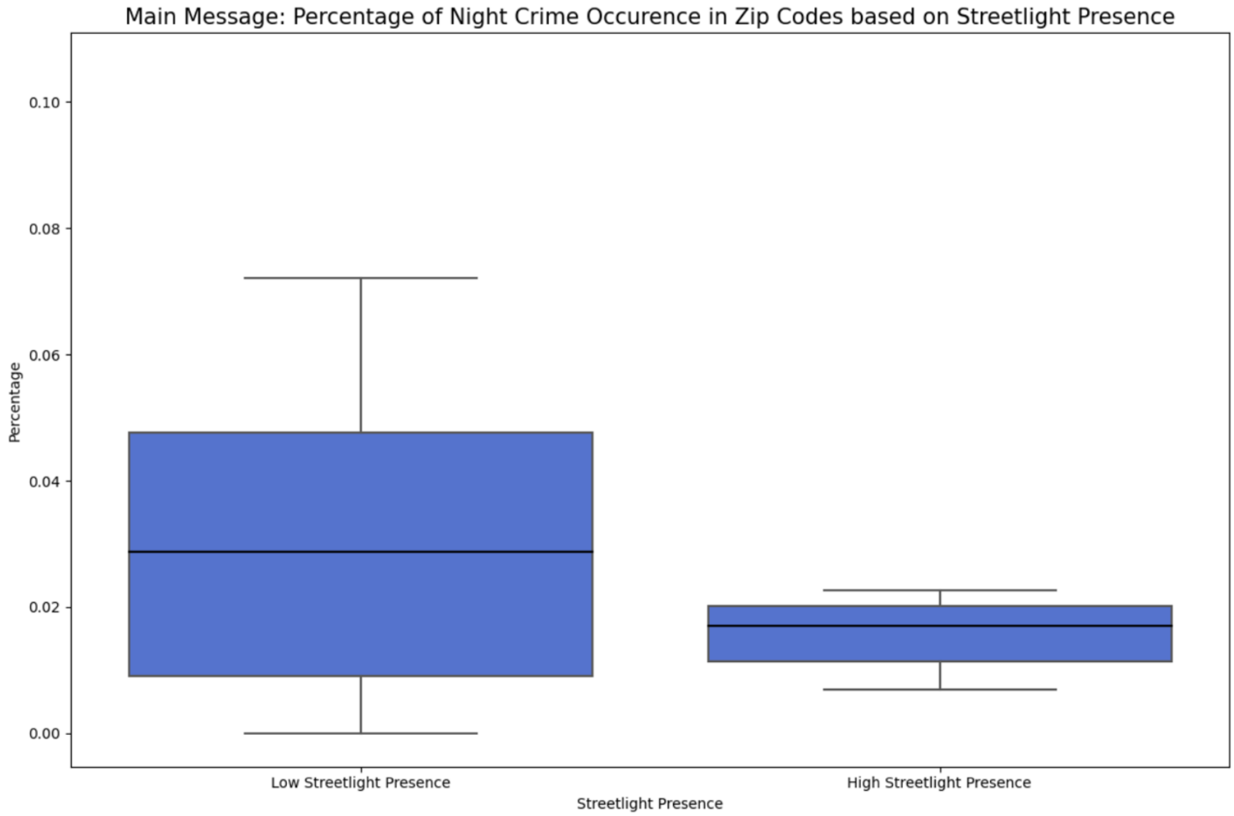


Figure 1.4 displays the number of crimes that involved shooting and the number of crimes that didn't at different hours throughout the day over the time interval of the dataset. This trend was observed consistently in every year.

There exists a clear difference between the trends of shooting and non-shooting crimes. While both counts start at a peak in the early hours of the morning and decline as dawn approaches, non-shooting crimes seem to follow a monotonous path during the day and into the evening and night. Shooting crimes, however, occur more frequently as the hour gets later (disappearing light). Crimes of different nature thus follow different trends.

Main Message



The Main Message figure shows clear differences between the frequency of crimes occurring in areas with low streetlight presence vs. high streetlight presence. The threshold that separates zip codes into these classifications is standardised.

Irrespective of income and other control variables zip codes that are categorised as having low streetlight presence experience more crime on average, which is congruent with the central hypothesis. This relationship is exactly what the research aims to explore through regressions and other visualisations. There are, of course, confounding variables that may be contributing to the drastic difference seen here. Attempting to identify and control for them will provide a more in-depth analysis. The relationship may thus be overstated here.

Maps

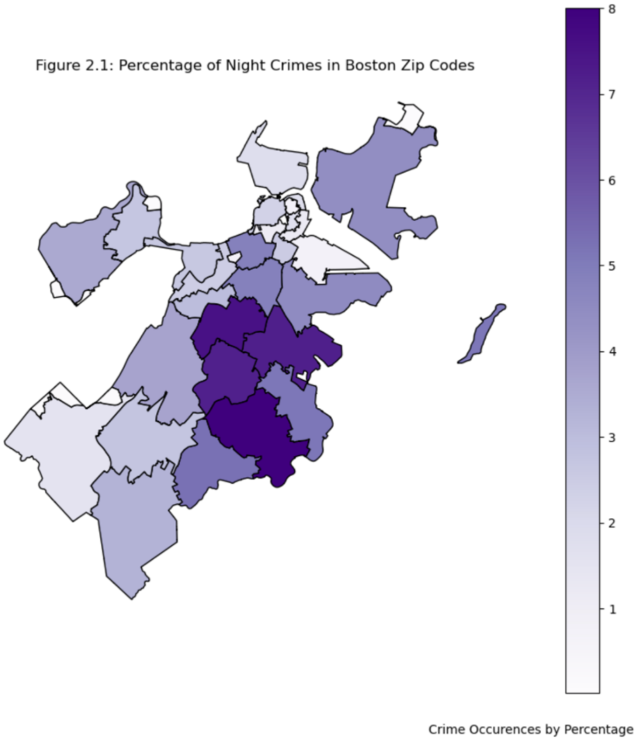


Figure 2.1: Percentage of Night Crimes in Boston Zip Codes

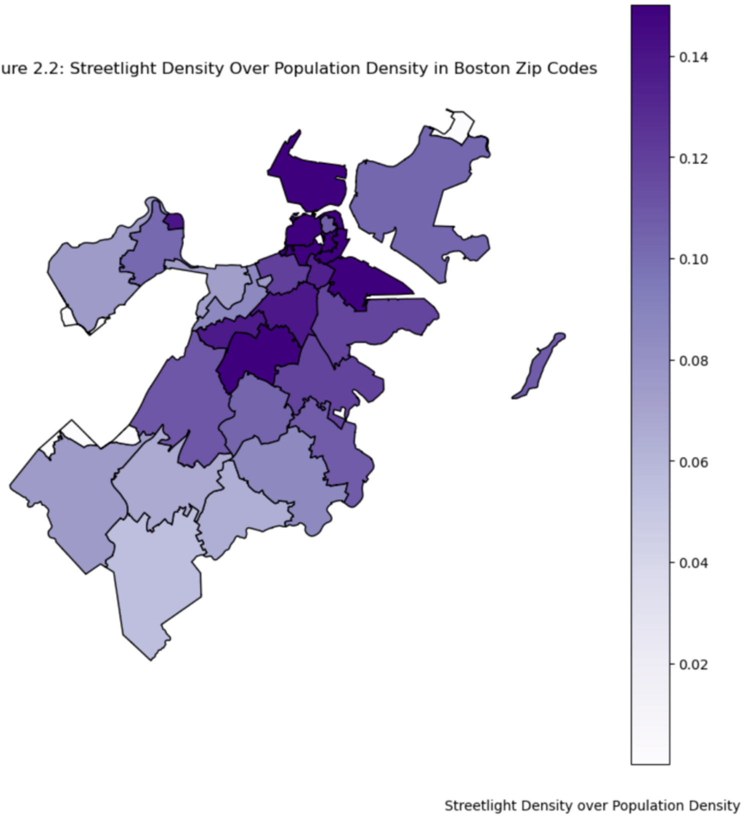


Figure 2.2: Streetlight Density Over Population Density in Boston Zip Codes

Figure 2.1 shows a map of Boston separated into zip codes. Each zip code is coloured purple based on the percentage of crimes in the dataset (subset for night crimes) that occur there. Most of the crime seems to occur in four zip codes. The purple gets less vivid the farther away the zip code is from this locality.

Figure 2.2 shows streetlight density over population density in order to meaningfully compare streetlight presence in different zip codes. Evidently, three out of the four zip codes where crime is concentrated (per Figure 2.1) are either moderately or weakly lit. It is also exhibited that well-lit zip codes experience the less crime. These observations support the main hypothesis of the research.

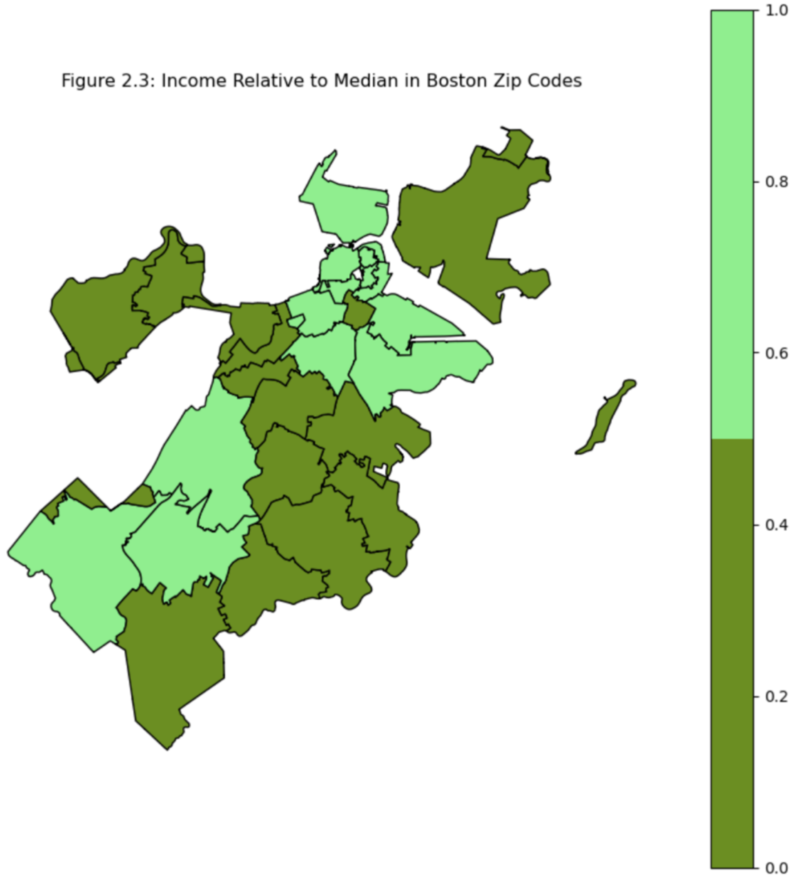
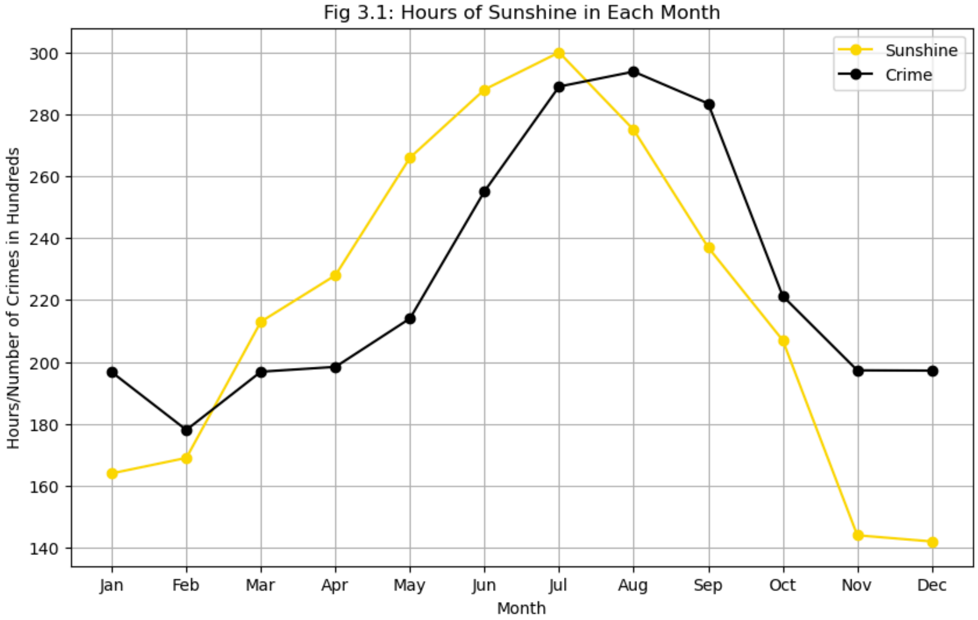


Figure 2.3: Income Relative to Median in Boston Zip Codes

Figure 2.3 shows high income vs. low-income zip codes. Classifications were made by finding the median income across Boston and comparing the average income in each zip code to this median. Higher than median was defined as high-income, while lower than median was defined as low-income. As hypothesised, higher income zip codes are comparatively more well-lit than low-income ones. Additionally, higher income zip codes experience less crime which may be indicative of the quality of life and security measures which prevent criminal activity.

It is important to note that a distinction has not been made with respect to whether zip codes are urban or residential areas. While this could potentially bias comparisons, accounting for population density attempts to act as a proxy since the difference between urban and residential areas would be the amount of people/traffic they experience.



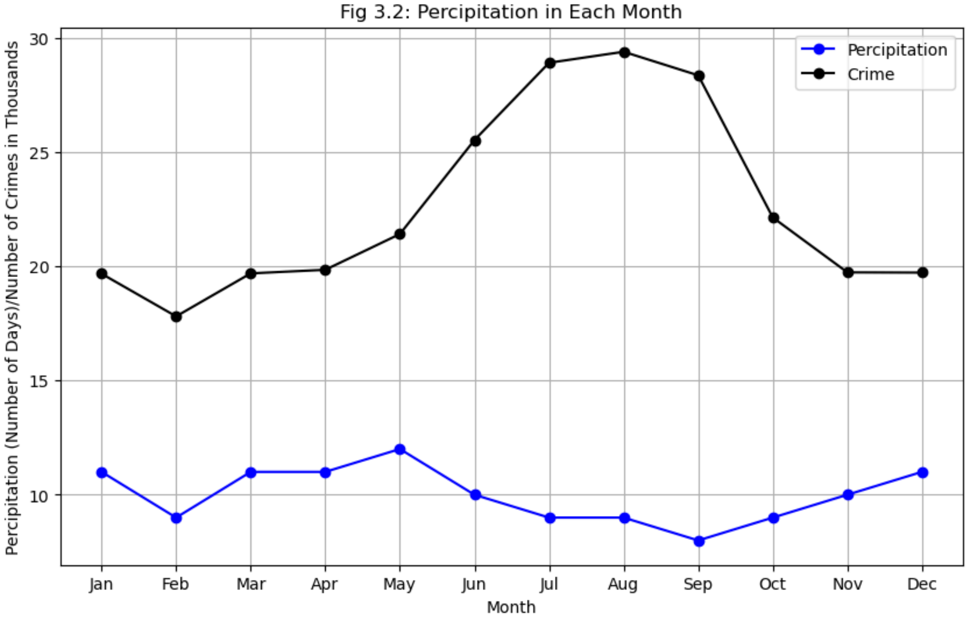


Figure 3.1 and 3.2 show hours of sunshine and precipitation in each month compared to crime counts in those months. Sunshine and precipitation data were scraped from the U.S Climate Data website.

It can be seen from Figure 3.1 that hours of sunshine and the frequency of crimes are positively related. Previous discussion has attributed this to human routine, police reporting, and victim realisation. Thus, figure 3.1 reinforces previous visualisation and allows greater focus to be placed on the impact of artificial lighting, which changes the game for those crimes committed at night.

Figure 3.2 shows that drier months experience more crime. This cushions the argument that weather conditions in colder and wetter months may be unsuitable for criminal activity. Thus, the fall in crime in this time period compared to summer is more greatly influenced by environmental factors than by light.

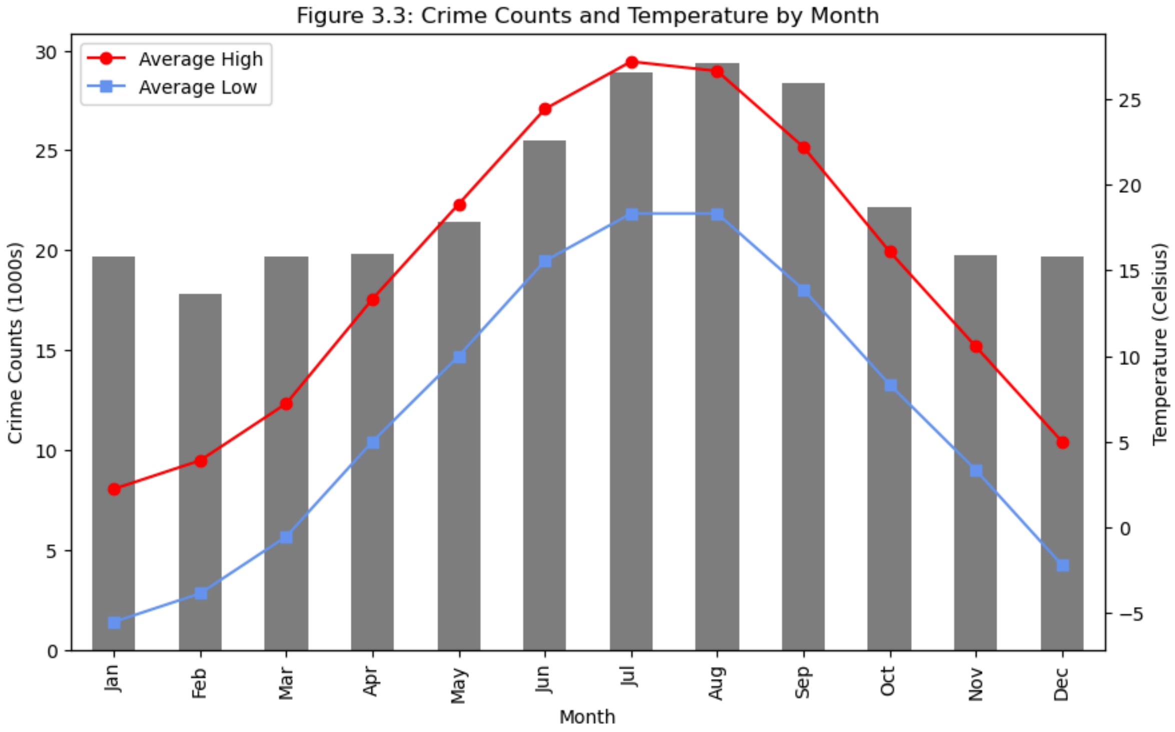


Figure 3.3: Crime Counts and Temperature by Month

Figure 3.3 brings in temperature data. This graph also coincides with previous discussions of the relationship between temperature (seasonal changes) and crime. Overall, crimes tend to follow the same pattern of occurrence as temperature. This provides credence to the idea that higher crime frequencies in summer and autumn can be attributed to favourable weather conditions rather than lighting.

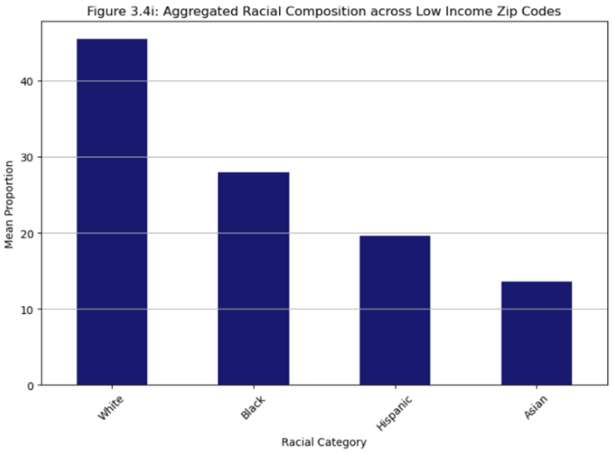


Figure 3.4i: Aggregated Racial Composition across Low Income Zip Codes

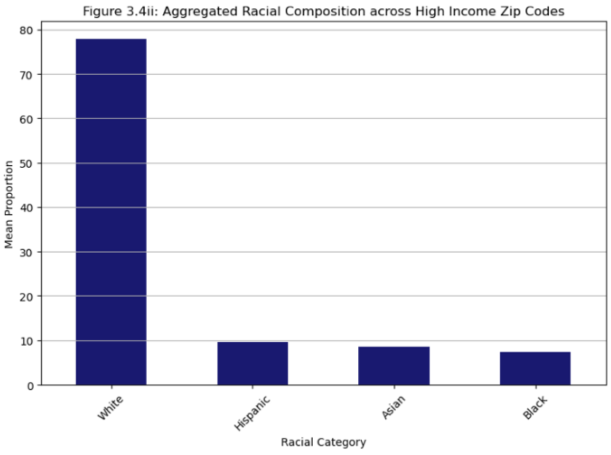
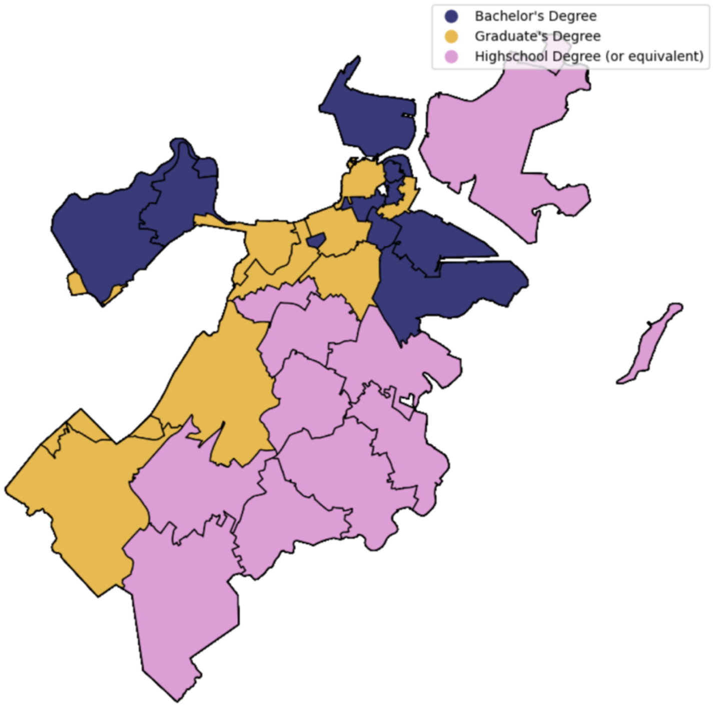


Figure 3.4ii: Aggregated Racial Composition across High Income Zip Codes

Boston is mostly composed of white people. This is true across zip codes however composition is different between low-income and high-income zip codes. This is an interesting observation because the frequencies of crime in each vary. As can be seen from Figures 3.4i and 3.4ii, the composition in low-income zip codes is far more distributed across racial categories compared to high-income zip codes which are made up of majority white people. This is a baseline difference that should be controlled for in regression analysis in order to make comparisons between zip codes comprable, particularly along lines of income.

Figure 3.5: Majority Education Levels in Boston Zipcodes



Consolidating Figure 3.6 with previous maps shows that lower educated areas experience more crime. In fact, all four zip codes experiencing the highest levels of crime have a maximum educational attainment of a high school degree (or equivalent). It is also evident that low-income zip codes more commonly don't achieve higher education. While there is variation of Bachelor's and Graduate degrees between high income zip codes, it is not unreasonable to say that zip codes with higher educational attainment in general are associated with less crime.

Results

Through the process of creating maps and data visualisations, the most likely relationship between the independent and dependent variable seemed linear. This is because a general negative relationship has been shown— zip codes with fewer streetlights relative to their population experience more crime. There does not seem to a be a threshold around which this relationship changes thus higher order polynomial relationships are unlikely. Moreover, current analysis doesn't present an intuitive reason for why there would be a non-linear relationship.

Existing theories have looked directly at the relationship between streetlight presence and crime frequencies. Some have also considered the time of day (the impact of Daylight Savings Time). Based off this, I will be including controls for hours of sunshine and precipitation in each month.

I will also be including demographic controls which, as discussed, differentiate zip codes from each other in terms of baseline characteristics. This will be a step towards isolating the impact of streetlights. Moreover, controls for income and unemployment will be included since these are two variables that also have a clear relationship to crime trends.

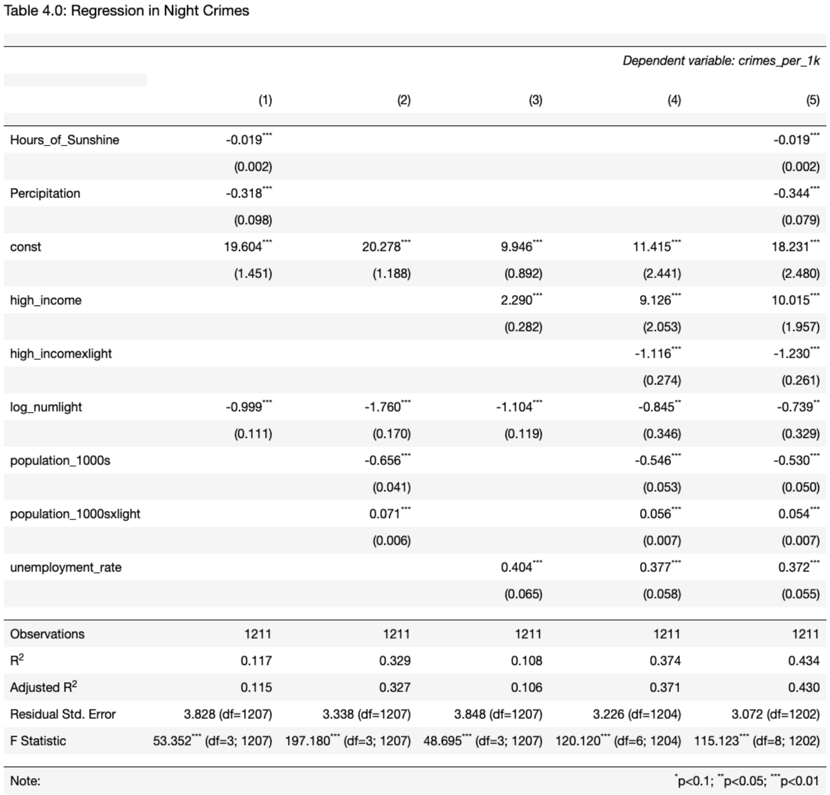
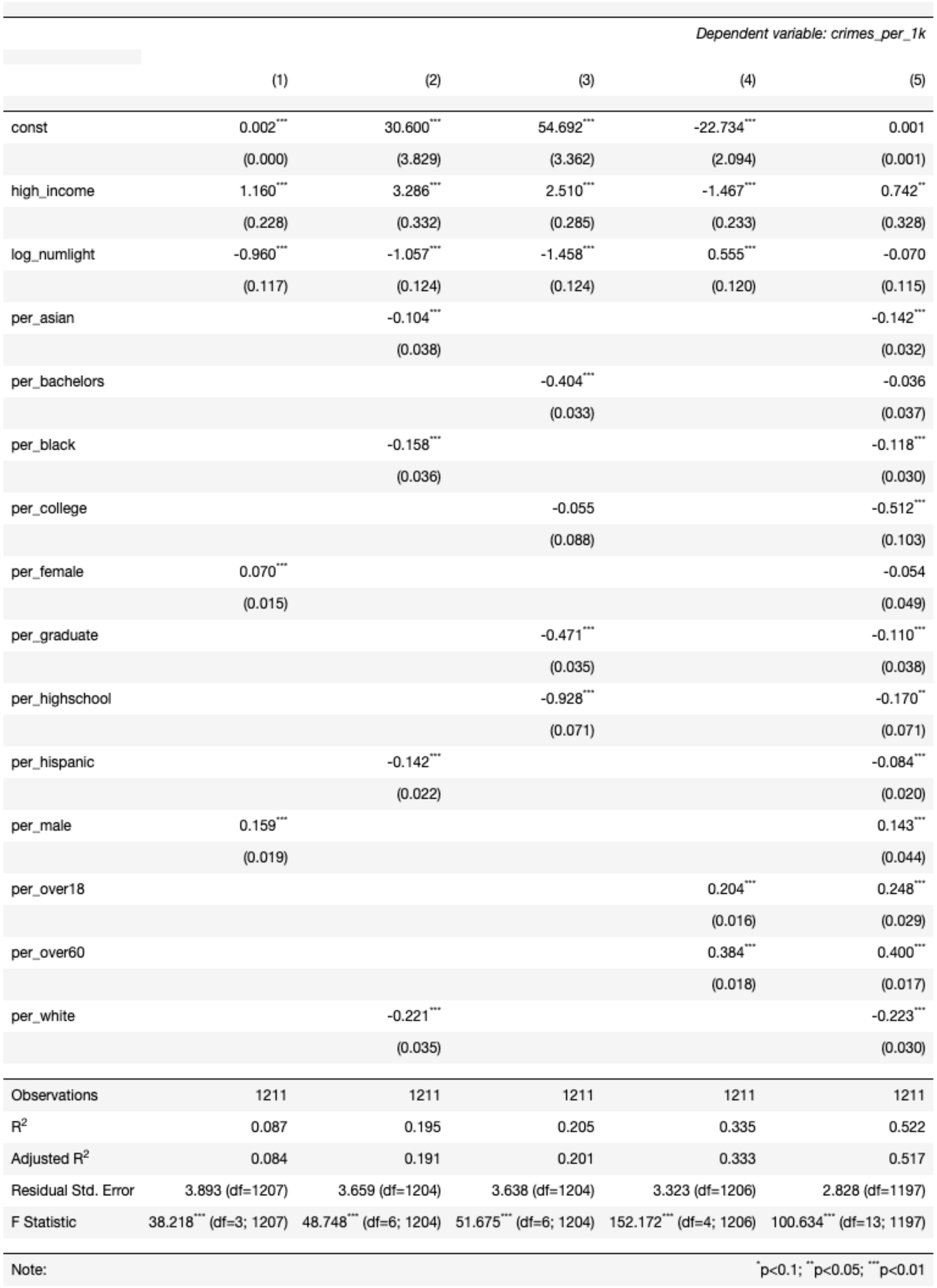


Table 4.0: Regression in Night Crimes

					Dependent variable: crimes_per_1k
	(1)	(2)	(3)	(4)	(5)
Hours_of_Sunshine	-0.019***				-0.019***
	(0.002)				(0.002)
Percipitation	-0.318***				-0.344***
	(0.098)				(0.079)
const	19.604***	20.278***	9.946***	11.415***	18.231***
	(1.451)	(1.188)	(0.892)	(2.441)	(2.480)
high_income			2.290***	9.126***	10.015***
			(0.282)	(2.053)	(1.957)
high_incomexlight				-1.116***	-1.230***
				(0.274)	(0.261)
log_numlight	-0.999***	-1.760***	-1.104***	-0.845**	-0.739**
	(0.111)	(0.170)	(0.119)	(0.346)	(0.329)
population_1000s		-0.656***		-0.546***	-0.530***
		(0.041)		(0.053)	(0.050)
population_1000sxlight		0.071***		0.056***	0.054***
		(0.006)		(0.007)	(0.007)
unemployment_rate			0.404***	0.377***	0.372***
			(0.065)	(0.058)	(0.055)
Observations	1211	1211	1211	1211	1211
R^2	0.117	0.329	0.108	0.374	0.434
Adjusted R^2	0.115	0.327	0.106	0.371	0.430
Residual Std. Error	3.828 (df=1207)	3.338 (df=1207)	3.848 (df=1207)	3.226 (df=1204)	3.072 (df=1202)
F Statistic	53.352*** (df=3; 1207)	197.180*** (df=3; 1207)	48.695*** (df=3; 1207)	120.120*** (df=6; 1204)	115.123*** (df=8; 1202)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 4.1: Regression in Night Crimes



	Dependent variable: crimes_per_1k				
	(1)	(2)	(3)	(4)	(5)
const	0.002***	30.600***	54.692***	-22.734***	0.001
	(0.000)	(3.829)	(3.362)	(2.094)	(0.001)
high_income	1.160***	3.286***	2.510***	-1.467***	0.742**
	(0.228)	(0.332)	(0.285)	(0.233)	(0.328)
log_numlight	-0.960***	-1.057***	-1.458***	0.555***	-0.070
	(0.117)	(0.124)	(0.124)	(0.120)	(0.115)
per_asian		-0.104***			-0.142***
		(0.038)			(0.032)
per_bachelors			-0.404***		-0.036
			(0.033)		(0.037)
per_black		-0.158***			-0.118***
		(0.036)			(0.030)
per_college			-0.055		-0.512***
			(0.088)		(0.103)
per_female	0.070***				-0.054
	(0.015)				(0.049)
per_graduate			-0.471***		-0.110***
			(0.035)		(0.038)
per_highschool			-0.928***		-0.170**
			(0.071)		(0.071)
per_hispanic		-0.142***			-0.084***
		(0.022)			(0.020)
per_male	0.159***				0.143***
	(0.019)				(0.044)
per_over18				0.204***	0.248***
				(0.016)	(0.029)
per_over60				0.384***	0.400***
				(0.018)	(0.017)
per_white		-0.221***			-0.223***
		(0.035)			(0.030)
Observations	1211	1211	1211	1211	1211
R^2	0.087	0.195	0.205	0.335	0.522
Adjusted R^2	0.084	0.191	0.201	0.333	0.517
Residual Std. Error	3.893 (df=1207)	3.659 (df=1204)	3.638 (df=1204)	3.323 (df=1206)	2.828 (df=1197)
F Statistic	38.218*** (df=3; 1207)	48.748*** (df=6; 1204)	51.675*** (df=6; 1204)	152.172*** (df=4; 1206)	100.634*** (df=13; 1197)
Note:					*p<0.1; **p<0.05; ***p<0.01

Regression 4.0

Model 1 regresses the logged number of streetlights, hours of sunshine, and precipitation on night crimes per 1000 people in a year/month/zip code combination. This was done first and foremost to see whether these weather and environmental characteristics would significantly contribute to the dependent variable. The coefficients are both statistically significant. While this is a subset for night crimes, the coefficient on Hours of Sunshine indicates that a month/year/zip code combination that experiences 100 more hours of sunshine is associated with 1.9 fewer night crimes per 1000 people on average, controlling for number of streetlights and precipitation. This relationship is reflected in the coefficient on precipitation as well which is unsurprising in accordance with Figures 3.1 and 3.2.

Model 2 regresses the logged number of streetlights, population in 1000s, and an interaction variable between the two on night crimes per 1000 people in a month/year/zip code. Intuitively, population would play a big role in crime frequencies between districts. More people in an area can mean more criminals but also more chances of being caught or seen. The negative coefficient indicates that, ceteris paribus, an increase of 1000 people in a zip code is associated with 0.6 fewer night crimes per 1000 people which is statistically significant but not economically notable. The positive coefficient on the interaction recalls discussions regarding the impact of crowds on visibility— where light exposes, people provide cover.

Model 3 regresses the logged number of streetlights, a dummy variable for high income zip codes, and the unemployment rate on night crimes per 1000 people in a year/month/zip code combination. Income is another characteristic that contributes to crime trends. This was seen in the various maps, where low-income districts were observed to experience most of the night crime in the dataset. The coefficient on unemployment rate is positive. Ceteris paribus, a 1% increase in unemployment is associated with 0.4 additional night crimes per 1000 people in a year/month/zip code combination on average. This is not economically significant, but its statistical significance and direction make it an important consideration.

Model 4 regresses the logged number of streetlights, dummy variable for high income zip codes, an interaction between the two on night crimes per 1000 people, population in 1000s and its interaction, as well as the unemployment rate on crimes per 1000 people in a year/month/zip code combination. This is the preferred specification and shall be discussed further.

Model 5 includes all controls in Table 4.0. The coefficient on the number of streetlights stays statistically significant and negative. All controls are also statistically significant, accounting for 43% of the variation in crimes per 1k people in a year/month/zip code combination.

Regression 4.1

Model 1 regresses the logged number of streetlights, income and the corresponding interaction, and sex demographics on night crimes per 1000 people in a year/month/zip code combination. This was done to see whether an increase in the population of males vs females in a zip code would have any relationship to crime. While neither coefficient is economically significant, the coefficient on female is statistically significant and negative. This indicates that when controlling for the proportion of males in the zip code population and the number of streetlights in the same area, a 10% increase in the proportion of females is related to approximately 1 less crime per 1000 people in that zip code on average.

Model 2 regresses the logged number of streetlights, income and the corresponding interaction, as well as race controls on night crimes per 1000 people in a year/month/zip code combination. Since high income and low-income districts have severely different distributions of race in their composition, these controls were thought relevant. While the interpretations of the coefficients aren't particularly relevant to the central hypothesis, the R^2 of 0.19 indicates that 19% of the variation in the independent variable can be explained by the variables in Model 2. This demonstrates the overall contribution of race controls.

Model 3 regresses the logged number of streetlights, income and the corresponding interaction, as well as controls for education on night crimes per 1000 people in a year/month/zip code combination. Education is another characteristic that was observed as a baseline difference between zip codes. Similar to race controls, the individual coefficients don't provide direct insight to the research question. However, the R^2 shows that 20% of the variation in the independent variable can be explained by the variables in Model 3. Thus, educational attainment controls are good observables to control for.

Model 4 regresses the logged number of streetlights, income and the corresponding interaction, and age controls on night crimes per 1000 people in a year/month/zip code combination. Age is also an important demographic factor. This model demonstrates its statistical significance. The positive coefficient on per_over60 indicates that ceteris paribus, year/month/zip code combinations with a 10% larger proportion of people over the age of 60 are associated with 3.8 additional night crimes per 1000 people on average. While this seems strange at first, it must be noted once again that the dataset concerns crime reports. It is unlikely that this result indicates that old people commit more crimes, rather that old people report them more often. This conclusion is intuitive and cannot actually be supported here but it's a reasonable angle to take when interpreting this model.

Model 5 includes all controls in Table 4.1. While the coefficient on the number of streetlights stays negative, it is no longer statistically significant. This indicates that the demographic make-up of a zip code is of greater importance in explaining criminal activity. These controls account for 52% of the variation in crimes per 1k people in a year/month/zip code combination, which is a true majority. Consolidating this with Table 4.0, it can be asserted that while streetlighting can have a significant effect on crime trends when controlling for income, unemployment, and population, demographic differences are the driver of criminal patterns. This may be a broader indication that baseline characteristics are correlated with streetlight allocation since racial and educational distributions are different between high-income and low-income zip codes and urban funding is likely related to existing zip code wealth.

Preferred Specification

$$nightcrimesper1k_i = \beta_0 + \beta_1 \log \_\widehat{numl}\imath ght_\imath \\ + \beta_2 h\imath gh\_\widehat{\imath ncome}_\imath + \beta_3 h\imath gh\_\widehat{\imath ncome}xl\imath ght_\imath + \beta_4 unemplo\widehat{yment}\_rate_\imath + \beta_5 popula\widehat{t\imath on}\_1000s_\imath \ \ \beta_6 populat\imath on\_\widehat{1000}sxl\imath ght_\imath + u_\imath$$

My preferred specification is Model 4 from Regression Table 4.0. Population is important because it contributes to the possible number of criminals in a zip code. It may also be a determinant of how the city chooses to allocate streetlighting. Similarly, income contributes to the allocation of lights as richer areas may warrant greater funding. Income also affects how people approach crime.

Specification 4 displays that controlling for the unemployment rate, population in thousands, income, and their respective interactions with the number of streetlights, a 10% increase in the number of streetlights in a zip code is associated with 8.5 fewer night crimes per thousand people in a year/month/zip code combination on average. This result is highly statistically and economically significant, supporting the central hypothesis of an inverse relationship between streetlighting and night crime.

To evaluate the regressions, take note of the R^2 and the F-statistic. The F-statistic for all models in both regression tables is high. As control variables are added the F-statistic increases, particularly if the controls are good explanatory variables for the independent variable. A large F-statistic is good because it shows that the variation observed in the regression can be attributed to more than just chance. The R^2 in each of the models varies but is not objectively low for any of them. My preferred specification has an R^2 of 0.374, indicating that 37.4% of the variation in night crimes per 1000 people in each zip code/year/month combination is explained by the variables in Model 4 (Table 4.0). This is good considering all of the data was purely observational.

All models in Table 4.0 show statistically significant negative coefficients on log_numlight which reinforces the central hypothesis that streetlight presence and crime occurrences are negatively related. Table 4.1 shows the same relationship with the exception of Model 4 which controls for age. It is unclear why this is occurring since age was not a particularly varying characteristic between zip codes. Since the relationship is highly statistically significant in the close neighbourhood around -1 for eight out of the ten regressions, it is not irrational to say that the discrepancy in Model 4 of Table 4.1 can be attributed to noise.

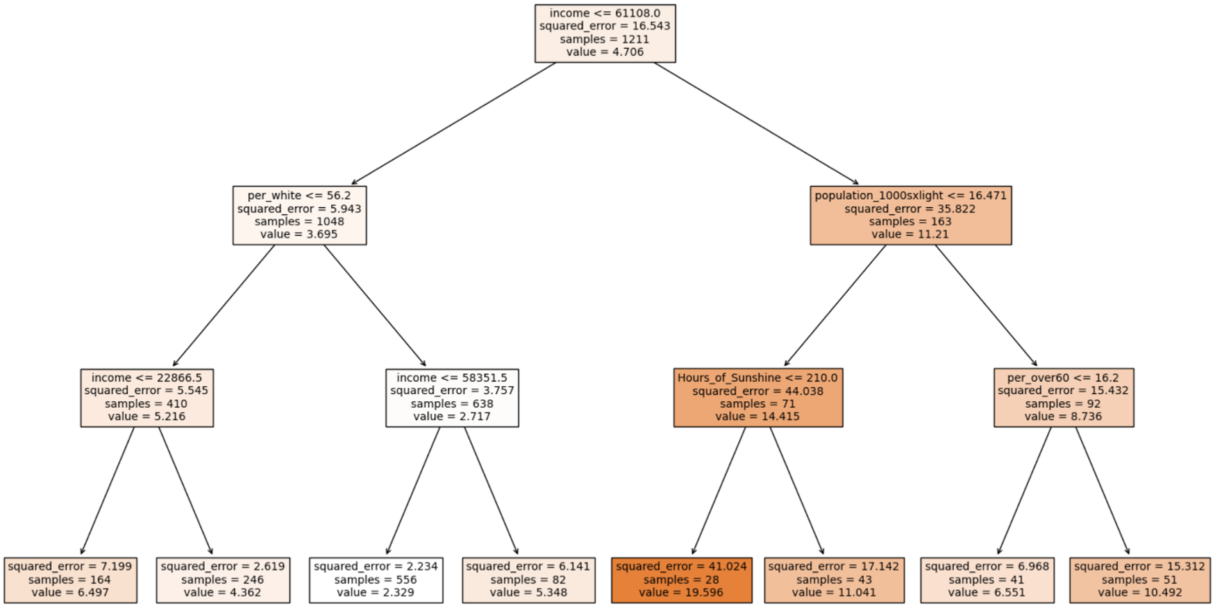
Machine-Learning

$$\boldsymbol{min_{j,s}}\left[\sum_{i:x_{i,j}\leq s,x_i\in R1}(\boldsymbol{crimesper1k_i}-\boldsymbol{cr\imath me\widehat{sper}1k_{R1}})^2+\sum_{i:x_{i,j}\leq s,x_i\in R2}(\boldsymbol{crimesper1k_i}-\boldsymbol{cr\imath me\widehat{sper}1k_{R2}})^2\right]$$

The objective function minimises the mean squared error between each feature and location. All the explanatory variables that this research uses will be stored in a rectangular space, R. The mean squared error (MSE) and root mean squared error (RMSE) describe how well the regression tree predicts the actual values of crimes per 1000 people for each year/month/zip code combination. The explanatory variables are iterated through and their MSEs minimised. The depth of a tree tells it when to stop minimising the squared errors. This is useful because the tree-building algorithm can continue until MSE is 0, perfectly predicting the independent variable using the dependent variables in the dataset. This, however, is not actually a good thing because inferences from that tree cannot be extrapolated to populations outside of the dataset as the algorithm has 'over-fit' its predictions.

Regularisation parameters are used to prevent over-fitting. Maximum tree depth is an imposition on decision-making algorithms. Rather than focusing on nodes, it focuses on the number of sub-levels, or depth. The regression tree will split nodes until the maximum depth is reached. This prevents overcomplexity (hinders the tree from becoming too complicated based on specific samples that exist in the dataset) but if the depth is set too low then the tree may under-fit. A process known as tuning is used to find the optimal depth for a tree. The depth here has been set to three because tuning is outside the scope of this analysis and higher depths don't make a huge difference in terms of minimising the MSE.

Figure 4.1: Regression Tree



The regression tree starts at the root node. The decision-tree iterated through the entire dataset (subset for night crimes), minimising squared errors to choose the best feature upon which to split the data. Income seems to be the most important explanatory variable for crimes per 1000 people in each year/month/zip code combination.

The second iteration follows the condition set in the root node: income < = 61,108. If this condition is true, the left branch is followed where per_white is the most important explanatory variable in this subset. If the condition is false, the right branch is followed where the interaction between population in thousands and the logged number of streetlights is the most important variable.

This implies that the previously seen racial composition differences between high income and low-income zip codes contribute notably to how crimes per 1000 people are observed. In low-income zip codes, race seems to be more important than in high income zip codes (disproportionately white) where the interaction between population and streetlighting takes the title of most explanatory.

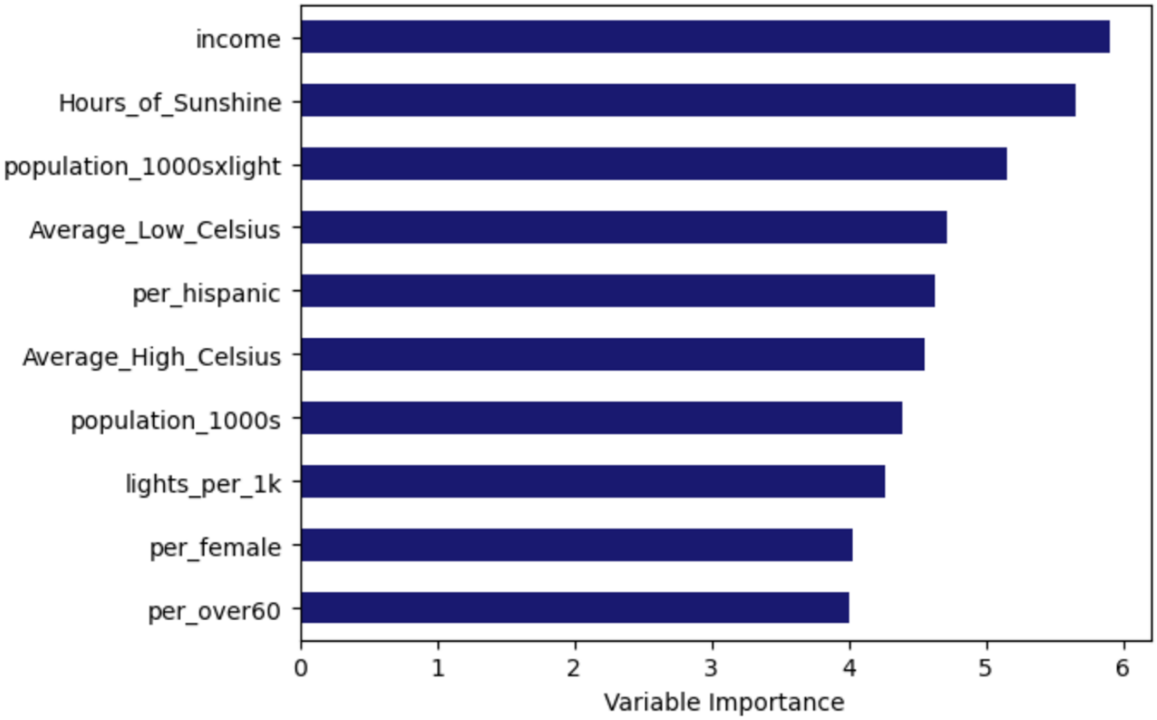
These nodes split one more time as the maximum depth of 3 is reached. On the left side, both nodes split back into income, indicating that income variation is an important variable even within low-income zip codes. On the right side, the split creates two nodes: hours of sunshine and proportion of population over 60.

The terminal nodes have subsets of the entire dataset following the previous leaves and their values predict the independent variable on average for the observations in that subset. For example, the left-most node says that in zip codes with an income of less than 61,108USD where the proportion of white people is less than 56.2% and the income in that subcategory is less that 22,866.5, the average crimes per 1000 people for each year/month/zip code combination in the sample of 164 is 6.5. Similar interpretations can be made for each of the terminal nodes.

It should be noted that number of streetlights on their own do not appear in the regression tree. While Table 4.0 and Table 4.1 mostly observed consistent statistically significant negative coefficients, the regression tree determined that income is the most important explanatory variable.

A RMSE of 1.9 showed that the difference between the regression tree's predicted value for crimes per 1000 people in a year/month/zip code and the actual value is 1.9 crimes per 1000 people on average. This is quite a small difference in the context of observational data, indicating the explanatory power of the regression tree.

Figure 4.2: Importance Matrix



The importance matrix ranks variables based on their efficiency at reducing the MSE when splitting a node. The above shows the top 10 most important variables. Among them are variables pertaining to environmental factors (hours of sunshine, average temperature highs and lows), population and demographic information, the interaction between population and logged number of streetlights, and income.

Income and a population interaction are in the top three which is unsurprising seeing as Model 4 in Regression Table 4.0 had a high F-statistic as well as a high R^2. Hours of sunshine also makes an appearance, indicating the importance of natural lighting and recalling the seasonal discrepancies between night and day crimes. Streetlights per 1000 people show up in the top 10 but not the top 3. Streetlighting may thus not be a driver of nighttime criminal activity, but the inverse relationship observed in Table 4.0 Model 4 indicates that it can be a mitigator.

Comparing the preferred specification to the regression tree and the importance matrix, the results are not drastically different. Income and population did turn out to be some of the most important explanatory variables. While my intuition led me to choose this specification, the most important variables may not have been as evident in a larger dataset with additional controls. The regression tree would solve this issue as it automatically minimised the mean squared errors to find the variables that best predict the independent variable. The regression tree also showed that the proportion of the population over the age of 60 is an important variable. This analysis doesn't analyse why zip codes with older populations are associated positively with number of night crimes reported per 1000 people, but this insight is interesting food for thought. It can also be a basis for thinking about future endeavours. Are older populations targeted more often? Do older people report more crime with respect to younger populations?

Conclusion

This project has endeavoured to explore how streetlighting affects night crimes. The central hypothesis was that zip codes with low streetlight presence would experience more crime, while zip codes with high streetlight presence would experience less. The possibility of this negative relationship was to be explored through a panel dataset spanning four years (2015-2016) from the Boston Police Department.

Summary statistics showed that high variations exist between zip codes regarding income, crime, and streetlight presence. They also displayed differences in crime trends over seasons and throughout the day. Interesting was the observation that night crimes tend to occur more often during winter and autumn while day crimes occur more frequently in spring and summer. This was attributed intuitively to social aspects like the facilitation of petty theft that the lively environments in summer and spring provide. It was also observed that crime increases in tandem with hours of sunshine and decreases with precipitation, which was also expected. Data visualisations showed that crimes of different nature may be associated with the high or low presence of light in different ways. These insights provided a well-rounded idea of how natural lighting was associated with crime trends, which was useful in developing intuitive explanations and thinking about confounding variables when conducting analysis on the subset of night crimes.

The various maps subset the original dataset for night crimes. This was where focus began to be placed on isolating the impact of artificial lighting in the zip codes. The maps reinforced the central hypothesis as zip codes experiencing the most crime were either weakly or moderately lit. These were compared to zip codes with the highest streetlight density/population density which experienced the least crime. Furthermore, mapping income compared to median separated the zip codes into high and low income. Zip codes were also categorized by their maximum educational attainment. This exhibited that the zip codes with the most crime and least light density relative to their population were low-income and high school educated, compared to high income and highly educated zip codes that were less concentrated with criminal activity.

Merging external data from the US Census Bureau allowed for the observation of baseline differences between the zip codes. These controls included race, sex, age, education, and unemployment rates. It was observed that low income and high-income zip codes have extremely different distributions of race in their composition. This guided thinking when it came to explaining the regression tree and models in the regression.

Running the regressions saw mostly consistent statistically significant negative coefficients of the logged number of streetlights. My preferred specification (Table 4.0: Model 4) indicated that controlling for unemployment rate, income, population, and their respective interactions with the number of streetlights, a 10% increase in the number of streetlights in a zip code is associated with 8.5 fewer night crimes per 1000 people on average. This is both a statistically and economically significant relationship. Model 4 was described as strong by performance statistics like the F-statistic. The variables of consideration in this Model made appearances in the importance matrix of a regression tree that was ran using a decision-making machine-learning algorithm.

In conclusion, this project has found a strong negative relationship between the presence of streetlighting, and night crimes based on a sample of crimes in Boston zip codes. While streetlighting is not the driver of crime trends at night, it can certainly be looked at as a deterrent. The limitations of this analysis lie in the fact that causal inference cannot be made. Future endeavours would aim to bring in data from other cities, categorise zip codes as urban or residential, and update streetlight information yearly. This would allow for a larger sample size and more variation in the data. Imposing the parallel trends assumption could thus potentially be justified if zip code fixed effects could efficiently be included. Causal inference through Difference-in-Difference could subsequently be pursued.

References

Analyze Boston. (2016). *Streetlight Locations.* Analyze Boston. https://data.boston.gov/dataset/streetlight-locations/resource/c2fcc1e3-c38f-44ad-a0cf-e5ea2a6585b5?inner_span=True

Boston Police Department. Uploaded by Ankurjain. (2019). *Crimes in Boston.* Kaggle. https://www.kaggle.com/datasets/ankkur13/boston-crime-data?resource=download

US Census Bureau. (2016). *American Community Survey 5-Year Data (2009-2022).* US Census Bureau.https://www.census.gov/data/developers/data-sets/acs-5year.html

US Census Bureau. (2016). *Index of Tiger GEO2016.* US Census Bureau. https://www2.census.gov/geo/tiger/GENZ2016/shp/

U.S Climate Data. (2024). *Climate - Boston, Massachusetts.* U.S Climate Data. https://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0601#google_vignette

Atkins, Stephen, Husain, Sohail, and Storey, Angele. (1991). The Influence of Street Lighting on Crime and Fear of Crime. *Crime Prevention Unit Paper* 28, 1-67.

Chalfin, Aaron, Hansen, B., Lerner, J. et al. (2022). Reducing Crime Through Environmental Design: Evidence from a Randomized Experiment of Street Lighting in New York City. *J Quant Criminol.* 38, 127-157.

Dominguez, Patricio, Kenzo, Asahi. (2023). Crime-Time: How Ambient Lighting Affects Crime. *Journal of Economic Geography.* 23, Issue 2, 299-317.

Pease, Ken. (1999). A Review of Street Lighting Evaluations: Crime Reduction Effects. *Criminal Prevention Studies* 10, 47-76.

Steinbach, Rebecca, Perkind, Chloe, Tompson, Lisa. et al. (2015). The Effect of Reduced Street Lighting on Road Casualties and Crime in England and Wales: Controlled interrupted Time Series Analysis. *J Epidemiol Community Health.* Published online: https://doi.org/10.1136/jech-2015-206012.

Xu, Yanqing et al. (2018). The impact of streetlights on spatial-temporal patterns of crime in Detroit, Michigan. *Cities.* 79, 45-52.