**Owen Galvin STAT E-100 Graduate Project**

**Introduction**

The original idea for the analysis in this project came from a quote I ran across in a [blog](hhttp://www.cambridgeday.com/2015/10/06/central-square-draws-most-concern-in-talk-of-drug-problems-worsening-everywhere/) where a Cambridge City Councillor was bemoaning the state of drug abuse in the Central Square area of Cambridge, saying that it was the worst he had ever seen it and ending with “People need to feel safe.” I have lived in Central Square for almost 5 years and as a resident it doesn’t strike me as having changed much one way or the other but I wanted to see how closely the “serious” crime rate tracked with the rate of what might be described as “Quality of Life” (QoL) crimes, those that generally wouldn’t affect me much as a passerby but would influence my perception of safety. Note: I avoid any formal interpretation in terms of the criminological “broken windows theory” but no doubt my general approach shares certain concepts. My designation of Serious vs. QoL crimes was mostly arbitrary, with part of the idea being that I wanted to concentrate on crimes that were more likely to be observed in public, or, if experienced, less likely to be reported to the police. This general idea leads to the largest assumption made in my approach, which is that the crimes in the dataset be considered an independent sample of the larger population of crimes-that-actually-occurred. No doubt there is a large corpus of research on types of crime and their likelihood for being reported to the police but that seemed well beyond the scope of this project. As a Central Square resident, some crimes such as “Drugs” seem relatively common and are unlikely to be reported to the police very often (only 517 in all of Cambridge over 5 years per the data seems rather low). On the other end of the scale, Homicides (11 total in original dataset) are generally going to be 100% reported. Another assumption regarding independence is that the time-based nature of the data, the effect of which hasn’t been covered in this course, won’t be a concern. I suspect the fact that crime reports are rolled up by year will obviate the need for special treatment but either way I wanted to sound a note of warning. Going back to the original quote, I initially tried to do some comparison of Central Square crime rates to non-Central rates, but the geographical nexus of what would generally be considered Central Square lies at the intersection of 4 neighborhoods and I only had population numbers at the neighborhood level.

**Data and Methods**

The primary data set is a list of crimes reported in the City of Cambridge, MA, more specifically

|  |
| --- |
| *“List of crime incidents featured in the Cambridge Police Department’s Annual Crime Reports and reported in the City of Cambridge over the last 5 years (2009-2013). Includes more than 40 crime types.**Certain crime types are excluded due to confidentiality and/or protection of privacy. “* <https://data.cambridgema.gov/Public-Safety/Crime-Reports/xuad-73uj> |

The dataset is available is available on the Cambridge Open Data [web site](http://data.cambridgema.gov) in multiple formats at the URL following the above quote.

*Dataset details:* Total of 36,497 observations spanning 5 years, 2009 through 2013, covering 7 variables:

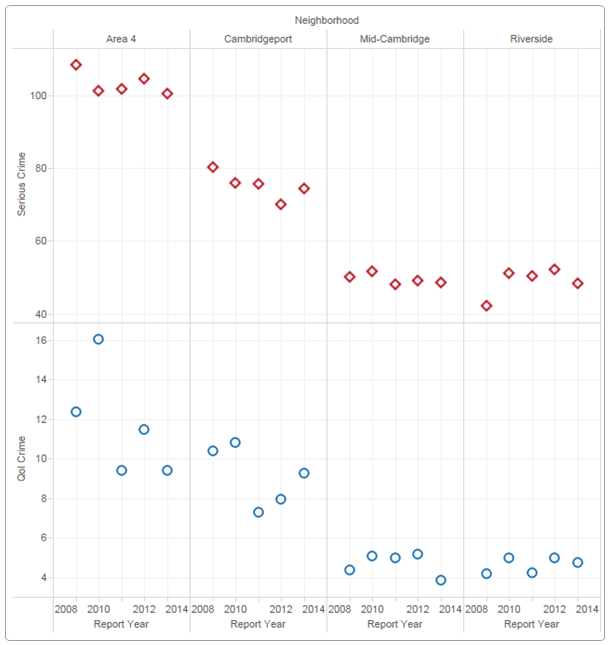
|  |  |
| --- | --- |
| **variable** | **Details** |
| FileNumber | Simple id in format of YYYY-######. |
| DateOfReport | Date and time of crime report. Parsed to year only for analysis purposes. |
| CrimeDateTime | Data and time, often a range, at which crime was reported to have happened. |
| Crime | One of 49 categories of crime. |
| ReportingArea | One of 116 sub-geographies in Cambridge. |
| Neighborhood | One of 13 neighborhoods as defined by the City of Cambridge. |
| Location | Rough addresses, generally within a block of actual reported location. |

The variables are generally nominal categorical values but the Date values can be considered ordinal if one is looking at crime over time. I then added a variable that categorized the crime values as being “Quality of Life” or not, with the 9 below granted the positive designation:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Annoying & Accosting | Disorderly | Drinking in Public | Flim Flam | Indecent Exposure |
| Harassment | Peeping & Spying | Stalking | Threats |  |

The data was imported into SQL for manipulation and combined with 2010 per-Neighborhood population data, drawn from a [demo graphic pdf](http://www.cambridgema.gov/~/media/Files/CDD/FactsandMaps/profiles/demo_profile_neighborhood_2013.ashx)  published by the City of Cambridge. This allowed for per-neighborhood/per-thousand-residents crimes rates, for both Serious and QoL crimes, resulting in 65 final observations for analysis – 13 neighborhoods x 5 years.

The first approach to comparing Serious vs. QoL crimes might simply be to visualize both rates for each neighborhood, essentially to eyeball before taking on a more analytic approach. The below chart generated in Tableau shows four of the neighborhoods, those that happen to border on Central Square. Serious crime rate is on top half and QoL on lower – simply looking at the height of each marker in a given year indicates there is likely some correlation, and good reason to proceed, but nothing too striking yet.



The primary analysis technique used to examine the correlation between Quality of Life crimes and Serious crimes will be a linear regression model. After that has been created I will run through some diagnostic checks to see how well model assumptions were met.

**Results**

We begin with a scatterplot with simple linear regression line, QolCrime as the predictor and SeriousCrime as the dependent variable:

|  |
| --- |
| > xyplot(crime$SeriousCrime ~ crime$QolCrime, data=crime, type=c('p','r'), main='Predicting Serious Crime from QoL Crime', xlab='Neighborhood Quality of Life Crime Rate (per 1000 pop.)', ylab='Neighborhood Serous Crime Rate (per 1000 pop.)') |

*Abbreviated model details:*

|  |
| --- |
| > model = lm(SeriousCrime ~ QolCrime, data=crime) > summary(model)  Residuals:  Min 1Q Median 3Q Max  -47.117 -10.321 -3.513 4.153 105.222  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 16.5501 5.4988 3.01 0.00376 \*\*  QolCrime 8.2067 0.7322 11.21 < 2e-16 \*\*\*  Multiple R-squared: 0.666, Adjusted R-squared: 0.6607  F-statistic: 125.6 on 1 and 63 DF, p-value: < 2.2e-16 |

Perform some diagnostic checks in regards to the regression assumptions:

1. Relationship between input and output variables are linear in the population
   1. The initial [scatterplot](#xyplot1) with regression line shows a linear relationship.
2. Standard Deviation of the outcome is same for all values of the input variable in the population
   1. Looks good overall, no fan-out effect going on, but the outliers are definitely a concern. Adding text labels to outliers shows they are mostly for Highlands neighborhood.

|  |  |
| --- | --- |
| > plot(crime$QolCrime, residuals(model), main='Residual Variability', xlab='QolCrime', ylab='Residuals') | > text(crime$QolCrime, residuals(model), labels=crime$valueDescription, pos=2) # + manual cleanup |

1. Outcome variable is normally distributed around the population regression line
   1. Those rightmost Highlands outliers aside, residuals look to be normally distributed.

|  |
| --- |
| > hist(residuals(model), breaks="FD", xlab="Residuals", main="Histogram of residuals")  > lines(-50:50, 560\*dnorm(x,0,sd(residuals(model))),col=2) |

Of the 13 neighborhoods for which crime rates were calculated, what is most notable about the one primarily responsible for the outliers, Highlands, is its relatively small size. The average neighborhood population in Cambridge for 2010 was around 8,000 but Highlands only had 832 people. Considering that associated five-year total number of Quality of Life is only 53, the presence of outliers is not surprising. It makes sense to create a new model that excludes Highlands entirely. Below is the linear regression plot using the new crime2 dataset – not outlier free but much better than the original model. (The graphic display of diagnostic regression checks is skipped below for space reasons but each of the checks result in fewer concerns about the new model’s viability.)

*A new Highlands-free model:*

|  |
| --- |
| > crime2 = subset(crime, crime$Neighborhood != 'Highlands')  > xyplot(SeriousCrime ~ QolCrime, data=crime2, type=c('p','r'), main='Predicting Serious Crime from QoL Crime', xlab='Neighborhood Quality of Life Crime Rate (per 1000 pop.)', ylab='Neighborhood Serous Crime Rate (per 1000 pop.)' |

One aspect to make note of, which can be seen by comparing the above regression line to that of the [earlier model](#xyplot1), is the somewhat less steep angle of the slope.

Finally we have the new model’s output, which will be analyzed as part of an overall hypothesis test:

|  |
| --- |
| > model2 = lm(SeriousCrime ~ QolCrime, data=crime2)  > summary(model2)  Call:  lm(formula = SeriousCrime ~ QolCrime, data = crime2)  Residuals:  Min 1Q Median 3Q Max  -31.591 -6.772 -0.612 5.249 35.741  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 19.5225 3.2397 6.026 1.24e-07 \*\*\*  QolCrime 7.0541 0.4792 14.722 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 11.53 on 58 degrees of freedom  Multiple R-squared: 0.7889, Adjusted R-squared: 0.7852  F-statistic: 216.7 on 1 and 58 DF, p-value: < 2.2e-16 |

**Model Interpretation**

Selected pieces of the R output can be examined to provide greater information about the model:

|  |  |
| --- | --- |
| Residuals:  Min 1Q Median 3Q Max  -31.591 -6.772 -0.612 5.249 35.741 | The 31.6 & 35.7 represent outliers, a residual plot on model2 shows 3 values near to 35 and only one at -31.6, either way they are from different neighborhoods, with no obvious reason to exclude. |
| Estimate  (Intercept) 19.5225  QolCrime 7.0541 | The intercept slope implies a 0 QolCrime rate would still result in 19.5 SeriousCrime rate, which isn’t particularly meaningful. The QolCrime slope implies each 1 point increase in QolCrime can be expected to result in a 7 point increase in SeriousCrime. The earlier model, with Highlands crime data included, had a steeper slope = 8.207. |
| Pr(>|t|)  (Intercept) 1.24e-07 \*\*\*  QolCrime < 2e-16 \*\*\* | Very small p-values along with the \*\*\* indicate that even with an alpha < = 0.001 these results would still be statistically significant. |
| Multiple R-squared: 0.7889 | The reported R-squared value is notably high, showing that QolCrime accounts for almost 80% of the variation of SeriousCrime in model2. |

**Confidence Interval**

|  |  |
| --- | --- |
| > confint(model2)  2.5 % 97.5 %  (Intercept) 13.037482 26.00761  QolCrime 6.094978 8.01324 | A confidence interval can also be created, which indicates 95% confidence that the true slope for Qolcrime in the population is between (6.094978, 8.01324) and the true intercept in the pop- ulation between (13.0375, 26.008). |

**Hypothesis Test and Conclusion**

The null hypothesis (H0) will be that the QolCrime coefficient, i.e. slope, is 0 while the alternative hypothesis (Ha) is that the slope > 0, since the original goal was to see if QolCrime positively corresponds with higher SeriousCrime. The significance level will be standard α = 0.05. The t-value reported by R = 14.722 and since this is simple linear regression with only one input variable, degrees of freedom = 58 (60 observations - 1 – 1). Those values taken together result in a critical value, for a one-sided t-test, of approximately 1.67, significantly lower than 14.722 and indicative of H0 rejection in favor of Ha. Even simpler would have been to examine the reported p-value, which gives the probability of getting a t-statistic greater than 14.722 (or less than -14.722) on the t-distribution. That p-value, displayed as < 2e-16, is very close to zero (dividing in half for a one-sided test would have no practical effect) and so also strongly indicates the null hypothesis should be rejected in favor of the alternative.

A more meaningful restatement of the above conclusion is that there is very little chance the true serious crime rate in Cambridge is not correlated with the rate of what I’ve termed “Quality of Life” crime. To put a glossy spin on it, if you are a resident of Cambridge and find that over time you feel less and less safe as you walk around your neighborhood, there is a very good chance you are indeed less safe. That statement disregards any mention of exactly who crime victims are likely to be and any number of other variables that a true criminological analysis would require. Probably the biggest caveat would be to emphasize the fact that this (very brief) analysis is entirely observational in nature and only attempts to measure existing correlation. It says nothing about a rise in QoL crimes actually causing an increase in the serious crime rate, i.e. along the lines of the broken windows theory where civil disorder is thought to indeed lead to an increase in the amount of serious crime.

Given the stated limits of my study, at least one area I would be interested in learning about if possible would be what I alluded to a few sentences ago, where a casual reading of violent crime reports in the media leads one to believe that, in Cambridge at least, homeless residents are a relatively large proportion of such victims. One wonders if the QoL crime perception (which my model assumes is experience equally by all residents) leads to a significantly different correlation with the Serious crime rate depending on whether or not the latter victims are homeless. Does whether or not “I” feel safe really have as strong an effect on my likelihood of experiencing some aspect of Serious crime vs. that of a homeless person. That probably isn’t possible with current data being collected by the City of Cambridge but one simple improvement on the existing approach would involve a better understanding of each original crime category to help with more accurately being able to rate it as QoL or not. The original data didn’t have much in the way of documentation and I didn’t spend significant time diving into the terms used to label the types of crime.