Does crime affect school attendance when crimes happen near schools?

Recent studies have suggested that there is higher high school absenteeism when the neighborhood of residence for students is more dangerous due to the dangerous commute to schools[1]. However, what if the school itself is located in a more dangerous neighborhood? I am interested in seeing if that would have any effect on absenteeism of those high schools. In this study, my hypothesis is that there is a negative relationship between safety of school neighborhoods and school attendance - high schools that are located in areas with more crime will tend to have lower attendance rates.

My dependent variable in this study is *perc_attend*: The attendance rates of every high school in New York City in the 2019 school year, taken from the NYC Department of Education's High School Directory dataset. Since the observations are the zip codes in NYC and many zip codes have multiple high schools, my final dependent variable is the average attendance rate of all schools for each zip code. Some zip codes do not have any high schools, so instead of treating those observations as missing values, I impute the data by taking the average attendance rate of all zip codes (the average of the dependent variable) and replace the missing values with that. To measure the main independent variable, count_crime, - amount of danger in every zip code, I take the crime count of every zip code from the NYPD Complaint data.

Figure 1 shows a choropleth map that visualizes the zip codes based on crime count. Zip codes with high crime counts are indicated by darker shades of green, and zip codes with lower counts of crime are indicated by white or lighter shades of green. The map shows that crime levels are higher in Upper Manhattan and Lower Manhattan, as well as in east Brooklyn, while lower crime areas are closer to the edges or near the borders of Queens and Staten Island that are

[1] Julia Burdick-Will, Marc Stein, Jeffrey Grigg. Danger on the Way to School: Exposure to Violent Crime, Public Transportation, and Absenteeism. *Sociological Science*, 2019; 6: 118 DOI: 10.15195/v6.a5

farther away from Manhattan. This demonstrates that higher crime seems to lessen as one moves further from Manhattan or Brooklyn.

The red centroids on top of the choropleth map in Figure 2 represents the average proportion of student attendance for all the schools within each zip code. Smaller centroids indicate lower school attendance averages, and larger centroids indicate higher school attendance averages. The higher rates of attendance do appear to have a bit of clustering as well. Areas of high attendance tend to be located within Manhattan, east Queens, and southeast Brooklyn. While there are in fact areas of high crime count with lower average attendance rates in larger areas like the Bronx and east Brooklyn/west Queens, there are also areas of higher crime count that overlap with schools with larger attendance rates, which goes against my hypothesis. Therefore, the suspected pattern between attendance and crime is not significant enough to conclude that there is a negative pattern which exists solely based on visualization.

Figure 1

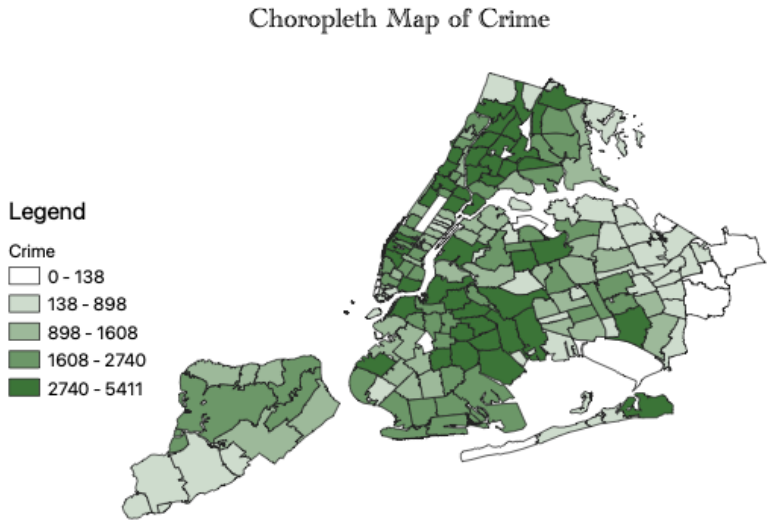
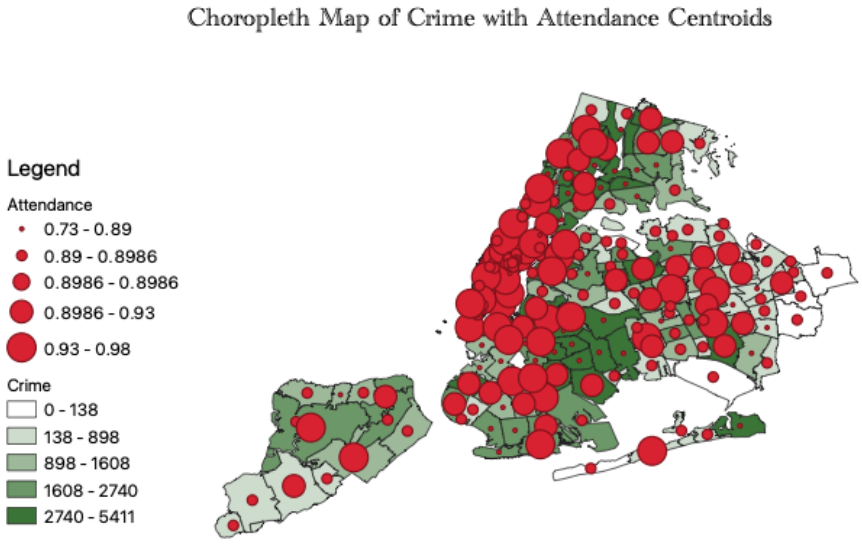


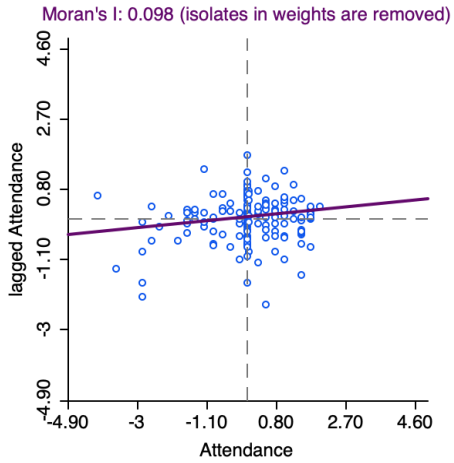
Figure 2



The Moran's I statistic for attendance is 0.099 with statistical significance at alpha < 0.05, which means that while there is conclusive evidence for a bit of spatial dependence, the extent of it is quite low due to the small magnitude of the correlation coefficient. The Moran's I scatter plot of attendance in Figure 3 also shows that places of high high school attendance have neighbors that also have high high school attendance, but the scatterplot line is quite flat, once again indicating that the magnitude of spatial dependence is not very large.

```
Moran's I Stat for attendance: 0.09860078641682433
Moran's I p-value: 0.015
```

Figure 3



For part of my analysis, I run a bivariate regression on three separate models using the default matrix (Queen's contiguity) as a spatial weight. For the first regression, I use the Ordinary Least Squares model to analyze the effect of crime count on attendance rate, and return the summary to interpret the results, displayed in Table 1. The average school attendance rate for every zip code is roughly 89%. The variable, count_crime, can be interpreted as: for every 1 increase in the number of crimes in a zip code, there is a decrease in proportion of overall school attendance by 0.0000052 on average. The probability value is 0.016, which suggests that the result is statistically significant at the 0.05 level. In addition to the beta coefficient, there are other values that are worth analyzing, such as r-squared. The r-squared value shows that crime count explains 2.7% of attendance.

Based on diagnostics for multicollinearity, we don't have to worry about the violation of the multicollinearity assumption, because the value is less than 20, at 2.68.

We use the Jarque-Bera test to check on the normality of errors. The value was 50.24 with high statistical significance, so we can reject the null hypothesis that the data is normally distributed.

The diagnostics for heteroskedasticity indicate that there is a presence of heteroskedasticity because the probability levels for the Breusch-Pagan test and Koenker-Bassett test results are very low, with values of 66.44 and 33.78 respectively. Thus, the variance of errors in this regression does in fact depend on crime count, which violates the Gauss-Markov Assumption of homoscedasticity. In this model we must reject the null hypothesis of homoscedasticity with confidence.

The Moran's I statistic for the residuals is 0.05 and is statistically insignificant because it's probability value 0.217. Therefore, we conclude that there is no spatial autocorrelation among the residuals that we need to be concerned about.

Table 1

```
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  Bivariate data of crime count on attendance rate
Dependent Variable  : perc_attend  Number of Observations:  214
Mean dependent var  :    0.898611  Number of Variables   :    2
S.D. dependent var  :   0.0411505  Degrees of Freedom    :  212

R-squared           :    0.027303  F-statistic           :     5.95081
Adjusted R-squared  :    0.022715  Prob(F-statistic)     :   0.0155326
Sum squared residual:    0.352485  Log likelihood        :      382.08
Sigma-square        :  0.00166267  Akaike info criterion :    -760.161
S.E. of regression  :   0.0407758  Schwarz criterion     :    -753.429
Sigma-square ML     :  0.00164713
S.E of regression ML:   0.0405848

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT      0.906457     0.00425601        212.983     0.00000
       count_crime  -5.19823e-06    2.13092e-06       -2.43943     0.01553
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   2.680743
TEST ON NORMALITY OF ERRORS
TEST                   DF           VALUE             PROB
Jarque-Bera             2            50.2492          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                   DF           VALUE             PROB
Breusch-Pagan test      1            66.4436          0.00000
Koenker-Bassett test    1            33.7832          0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : Queen's Contiguity
   (row-standardized weights)
TEST                          MI/DF        VALUE          PROB
Moran's I (error)             0.0533        1.2344        0.21706
Lagrange Multiplier (lag)       1           0.5716        0.44963
Robust LM (lag)                 1           0.1041        0.74696
Lagrange Multiplier (error)     1           1.1632        0.28080
Robust LM (error)               1           0.6957        0.40423
Lagrange Multiplier (SARMA)     2           1.2673        0.53065
============================== END OF REPORT ================================
```

The spatial lag model displayed similar results to that of the OLS model. The beta coefficient is a very slightly lower value in the spatial lag compared to the one in the OLS. It shows that for every increase in crime count per zip code, the attendance proportion decreases by an average of 0.00000511. This result is statistically significant once again at the 0.05 level because the p-value is 0.015. This model also includes the spatial lag term of attendance - *W_perc_attend* - as an additional indicator. This variable has a positive value of 0.030 and a probability value of 0.456. This shows that the average influence on observations from its

neighboring observations is 0.030, but because it is statistically insignificant, we conclude that there is no clear evidence of a spatial diffusion process. Other things to note is that the r-squared value is slightly higher than that of the OLS model by 3 percent, and the log likelihood is just slightly greater than that of the OLS by 0.28, suggesting that the performance of both models were somewhat similar.

Table 2

```
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Bivariate data of crime count on attendance rate
Spatial Weight      : Queen's Contiguity Matrix
Dependent Variable  : perc_attend  Number of Observations:  214
Mean dependent var  :    0.898611  Number of Variables   :    3
S.D. dependent var  :   0.0411505  Degrees of Freedom    :  211
Lag coeff.   (Rho)  :   0.0303374

R-squared           :    0.030066  Log likelihood        :     382.362
Sq. Correlation     : -            Akaike info criterion :    -758.724
Sigma-square        :  0.00164245  Schwarz criterion     :    -748.626
S.E of regression   :   0.0405271

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
     W_perc_attend     0.0303374      0.0406593       0.746137     0.45558
          CONSTANT      0.879145      0.0364929        24.0909     0.00000
       count_crime  -5.11801e-06    2.11848e-06       -2.41589     0.01570
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      1        64.3185     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : hw7nycw
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1         0.5634     0.45290
============================== END OF REPORT ================================
```

The spatial error model results in Table 3 displays similar results to what we observe in both the spatial lag and OLS models. The count_crime coefficient can be interpreted as: For every one increase in crime count, the average attendance proportion by zip codes decreases by 0.0000045, with a statistical significance at the 0.05 level. The coefficient value this time is lower than both the spatial lag and OLS models and the probability value of 0.00000437 is slightly higher than both of those models, but these differences are marginal and the overall results remain much the same. In this model, we include the coefficient of spatially correlated errors, referred to in the output as *LAMBDA*. This indicator has a positive effect of 0.105, but is

not statistically significant because the p-value is higher than 0.05 at 0.253. The p-value demonstrates that there is not a significant spatial dependent problem detected. Compared to the r-squared value and log likelihood values with the OLS model, the model improvement was negligible.

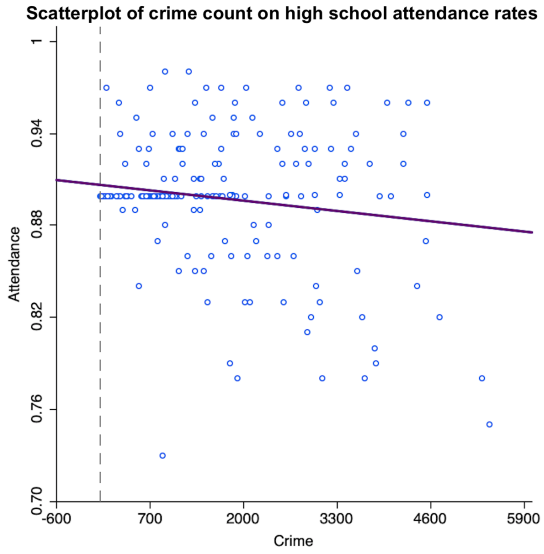
Table 3

```
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Bivariate data of crime count on attendance rate
Spatial Weight      : Queen's Contiguity Matrix
Dependent Variable  : perc_attend  Number of Observations:  214
Mean dependent var  :    0.898611  Number of Variables   :    2
S.D. dependent var  :    0.041150  Degrees of Freedom    :  212
Lag coeff. (Lambda) :    0.105308
R-squared           :    0.035279  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :  382.683685
Sigma-square        :  0.00163362  Akaike info criterion :    -761.367
S.E of regression   :   0.0404181  Schwarz criterion     :    -754.635

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
          CONSTANT      0.904989     0.00449943        201.134     0.00000
       count_crime  -4.45497e-06     2.2105e-06       -2.01537     0.04387
            LAMBDA      0.105308      0.0914614         1.1514     0.24957
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      1        61.6823     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : hw7nycw
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1         1.2064     0.27205
============================== END OF REPORT ================================
```

The results of all three bivariate models met my expectations in regards to my hypothesis. The results have consistently suggested a negative relationship between attendance of high schools and level of danger in high school areas within NYC. However, the magnitude of this effect is very small and close to zero, which suggests that the effect is marginal, and other factors may be far better predictors of attendance that were not included in this bivariate regression.

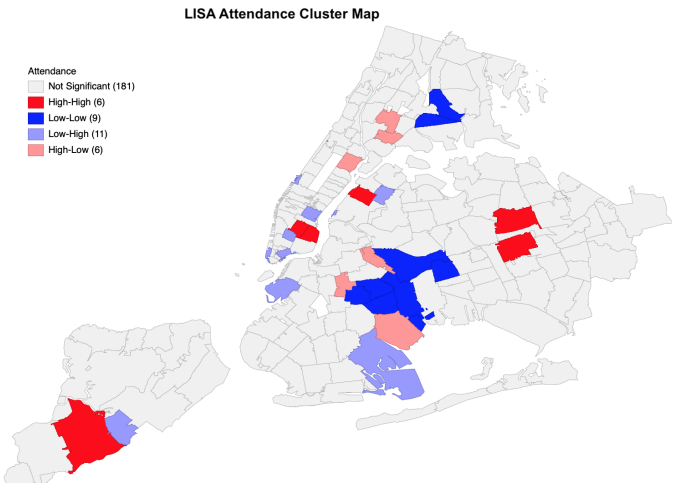
The scatterplot of crime on high school attendance visualizes the relationship between the two variables. The slope of the line shows that the effect goes in the negative direction that follows my hypothesis, but is not very steep. This indicates that the magnitude of the effect is not very large.

Figure 4



In addition, I created two separate Local Indicators of Spatial Association (LISA) cluster maps: one for crime, and one for attendance. The LISA cluster map for attendance in Figure 5 shows that there are a few clusters where high school attendance rates and its neighboring areas are high - namely, in the south of Staten Island, central Queens, and some smaller clusters on the east side of Manhattan. Some low-low clusters for areas of low school attendance are located in east Brooklyn and the Bronx. Lastly, it is worth noting that there are some areas that are spatial outliers and indicate no spatial clustering in small areas along east Manhattan and east Brooklyn.

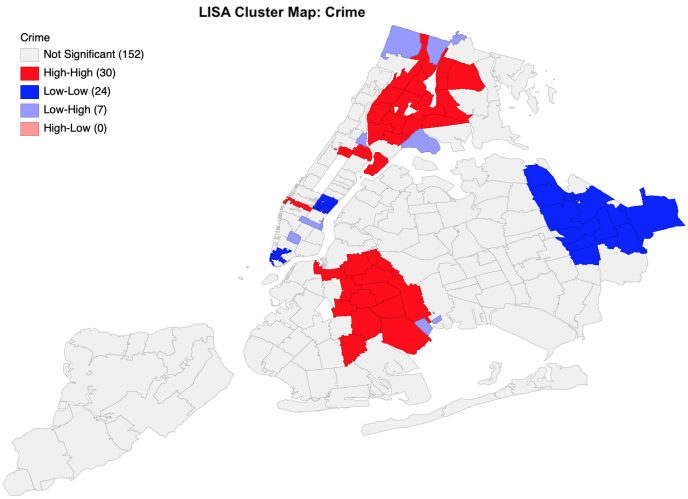
Figure 5



The LISA cluster map for crime in Figure 6 shows that there are two large clusters of high crime count: One in the east side of Brooklyn and the other located in the Bronx (upper Manhattan). Conversely, a large cluster of low crime is located in east Queens.

There are two noticeable areas where clusters of high crime and low attendance overlap, specifically east Brooklyn and a small portion of the Bronx. Besides that, there is not a significant pattern between the two LISA cluster maps.

Figure 6



To determine the final model between OLS, spatial lag, and spatial error, I used Anselin's diagnostic significance map. This method claims that if the Lagrange Multiplier (LM) and robust LM values for spatial and lag are statistically insignificant, then we do not need to use either of those models to control for spatial dependence. The spatial dependence diagnostics under the OLS model regression results displayed in Table 1 shows that the LM lag and the LM errors are all statistically insignificant because the probability values are far greater than 0.05. The p-values for the robust LM for lag and robust LM for error are also lower than 0.05. This shows that the OLS model is sufficient enough as the final model because there is not a detectable spatial dependence issue.

In my final model, I include three additional independent variables. I suspect that these variables - proportion of students who enrolled in a college or other post-graduate program within four years of graduating (*perc_college)*, school graduation rate (*perc_graduated*), and the proportion of students who feel safe when walking within schools (*perc_student_safe*) - may also contribute to school attendance in a different way. All of these new variables are proportions between the values of 0 and 1 like the dependent variable for attendance. The reason I added college enrollment and the graduation rate in particular is because some highly-ranked schools are also located within considerably dangerous areas of NYC. High-performing students who got accepted and enrolled into such high schools may be inclined to attend and not skip classes, regardless of the fact that it is located in an unsafe area. Due high school rankings being based partially on college preparation and graduation rate, I added these variables to analyze their individual effects on attendance as well as control for them in order to show the true (or at least more true) effect of the main independent variable, crime count, on attendance. I expect that all three control variables will have a positive effect on attendance.

For the results of my final model, the coefficient for crime count can be interpreted as: for every additional count of crime, the attendance proportion goes down by 0.0000013 on average. While this effect is consistent with my hypothesis, it is much lower than the effect from any of the bivariate regression models and is not statistically significant because the p-value is 0.26. As for the effects of the control independent variables, they are all statistically significant. For every increase in the perceived safety by 1 percent in each zip code, the attendance rate increases by 8 percentage points on average with high statistical significance. For every increase in the percentage of students enrolled in college or other post-high school graduate programs, the attendance rate increases by an average of 22 percentage points with high significance. Finally,

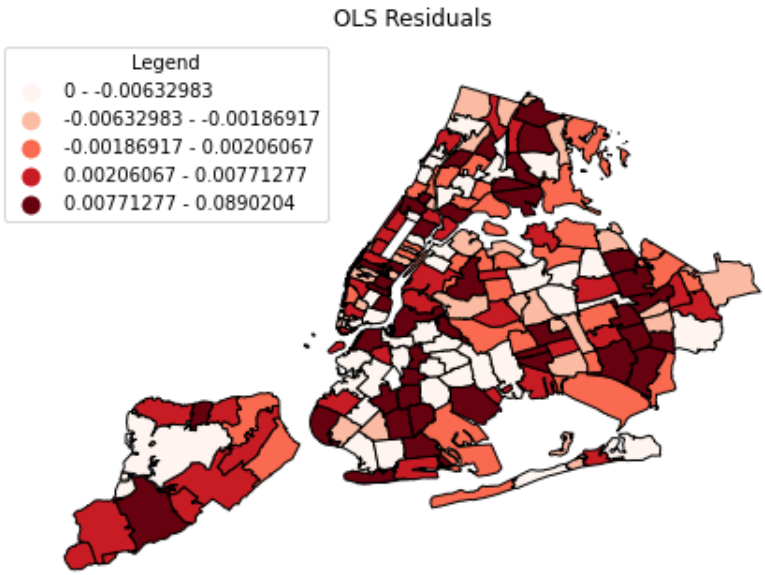
for every increase in percent of students who graduated in each zip code, the average attendance rate for all zip codes decreases by 2.5 percentage points with statistical significance at alpha = 0.05. Based on these findings, college enrollment and perceived safety have a positive correlation with high school attendance rates, with the latter having a much stronger correlation with attendance. While the perceived safety and college enrollment effects met my expectations, the graduation effect on attendance rates moved the opposite direction to what I had anticipated. However, the negative effect of graduation rates on attendance was quite small and not as significant in magnitude as the other two control variables. The new r-squared is very high compared to that of the original bivariate OLS model - adding three additional factors now increased the variance explained to 0.765, and the log likelihood is much higher at 534. Therefore, my model improved significantly from the bivariate OLS.

Table 4

```
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : MULTIVARIATE dataset of crime on attendance
Dependent Variable  : perc_attend  Number of Observations:  214
Mean dependent var  :    0.898611  Number of Variables   :    5
S.D. dependent var  :   0.0411505  Degrees of Freedom    :  209

R-squared           :    0.765354  F-statistic           :     170.426
Adjusted R-squared  :    0.760863  Prob(F-statistic)     :           0
Sum squared residual:   0.0850309  Log likelihood        :     534.234
Sigma-square        : 0.000406847  Akaike info criterion :    -1058.47
S.E. of regression  :   0.0201704  Schwarz criterion     :    -1041.64
Sigma-square ML     : 0.000397341
S.E of regression ML:   0.0199334

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT      0.697575      0.0233645        29.8561     0.00000
       count_crime  -1.29569e-06    1.14937e-06       -1.1273     0.26091
 perc_student_safe     0.0826622      0.0289914        2.85126     0.00479
      perc_college      0.224638      0.0116929        19.2115     0.00000
    perc_graduated    -0.0253502      0.0105804       -2.39596     0.01746
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   52.374387
TEST ON NORMALITY OF ERRORS
TEST                 DF           VALUE             PROB
Jarque-Bera           2           300.5369          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                 DF           VALUE             PROB
Breusch-Pagan test    4           124.3533          0.00000
Koenker-Bassett test  4            32.2477          0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : Queen's Contiguity
   (row-standardized weights)
TEST                         MI/DF        VALUE          PROB
Moran's I (error)           0.0673        1.5348        0.12483
Lagrange Multiplier (lag)      1          0.1747        0.67600
Robust LM (lag)                1          0.0158        0.89989
Lagrange Multiplier (error)    1          1.8551        0.17319
Robust LM (error)              1          1.6963        0.19277
Lagrange Multiplier (SARMA)    2          1.8709        0.39240
============================== END OF REPORT ==============================
```

In this final model's regression diagnostics, there is detection of multicollinearity because the multicollinearity condition number is greater than 20. Furthermore, the normality of errors and homoscedasticity assumptions are violated because in those diagnostic tests the probability value is lower than 0.001. However, there is still no detection of spatial dependence because the probability values for Moran's I error, and LM and Robust LM models, are all much greater than 0.05. In addition, the residuals map in Figure 3 shows that while there are a few clusters with high residuals or low residuals, they are for the most part, random.

Figure 7



Based on my overall findings, the alternative hypothesis that there is a negative correlation between crime count and attendance is rejected once adding in other factors that potentially account for attendance. This is based on my consistent results from my final model and spatial visualizations. My results from all three bivariate regression models, shows a small effect of crime on high school attendance. In addition, the final model that included the control variables was not statistically significant when controlling for college enrollment, perceived student safety within the school, and graduation rates. On the contrary, the control variables displayed a statistically significant positive correlation with school attendance when controlling

for crime, with graduation rates having the more significant effect on attendance based on the correlation coefficient magnitudes.