4031 Project Final Report

Homicide Rate Case Study

Team 24

Eric Gu, Ryan Reed, Michael Wang, Kexin Zhang

Table of Contents

I. Introduction

Of the many types of violent crimes committed in the United States every year, homicide and non-negligent manslaughter are among the most serious offenses. This category makes up a significant fraction of total nationwide violent crime, and is especially pertinent to the area surrounding Georgia Tech. Atlanta, which was named one of America's top 25 murder capitals in 2015, had a murder rate of 20.2 per 100,000 individuals, more than four times the nationwide average at the time.

Increased coverage of homicides, whether it be through traditional news outlets or through social media, has made it clear now more than ever the prevalence of homicide in America. While most attempts to generalize motives and link causality for homicide will be fruitless since motives are often ambiguous and impossible to test, our team believes that certain environmental factors, such as demographics or climate, may correlate with and serve as predictors for homicide rate. These environmental factors will be explained in greater detail in the Data Description section of our report.

II. Problem Statement

II.A. Question

To what extent can we predict the number of homicides in a city given quantitative and qualitative characteristics of that city?

II.B. Variable Summary

Dependent Variable	Independent Variables
Homicides per 100,000 people	Unemployment rate
	Population Density
	Average Temperature
	Precipitation
	Rate of Gun Ownership in State
	Median Household Income
	Census Region of the United States
	Percent of Population Identifying as Non-white
	Cost of Living
	High School Diploma Attainment Rate

Figure 1: Summary of Variables

II.C. Constraints and Assumptions

1. Some of the data was not available for all cities in the United States. Therefore, our team was biased towards relatively larger and more well known cities since they had complete data more often. In spite of this constraint, our team assumed the data contained a random distribution of cities across the United States.

2. Because finding cities that fit all the requirements was rare for most years, we chose to collect data for the year 2010 since that was a census year and variables were more accessible. We assumed that 2010 was a representative year for homicide rate and not an outlier.
3. For some of our independent variables, we could not get an exact measure since the statistic was not commonly recorded. Because of this, we used proxy statistics, such as median rent as a proxy for cost of living, after seeing this historically used in literature.

II.D. Goals

1. To identify which variables listed in Figure 1 are significant in predicting homicide rate
2. To quantify the extent to which the variability in homicide rate can be explained through variation in environmental variables
3. To greater understand homicide rate and ultimately improve the quality of life in US cities

II.E. Deliverables

1. A multiple linear regression model that predicts homicide rate given a combination of explanatory variables
2. Detailed analysis of methods, residual and multicollinearity tests
3. Recommendations for further analysis

III.Data Description

City	Murder rate per 100,000	State	Population Density	Gun Ownership in State	…	HS Grad Rate
Abilene	3.4	TX	408.5968586	6.973192644	…	81
Albuquerque	7.7	NM	1111.941332	9.818306785	…	85.2
Allentown City	7.6	PA	2538.322581	5.616794106	…	86.2
Ann Arbor	0	MI	1532.813131	5.892986107	…	92.9
…	…	…	…	…	…	…
Wichita	4.2	KS	902.4498466	7.224117009	…	87.9

Figure 2: Excerpt from Data Tables (100 Rows)

We focused on relatively sizable cities in the United States in 2010. The most recent census was conducted in 2010, and thus, the most accurate, reliable information is from that year. We also limited our scope to cities with populations of at least 100,000, as the FBI's dataset on city homicide rates includes only cities of at least 100,000 people. From that list of cities, we randomly selected one hundred cities and found the appropriate data for each.

The cities' homicide rates in 2010 serves as our dependent variable. The FBI defines homicide rate as the number of homicides and nonnegligent manslaughters per 100,000 people. The FBI's Uniform Crime Reporting Statistics program provide annual statistics for homicide rates for each city's police bureau. Typically, each city has its own police bureau, such as Atlanta's Atlanta Police Department, which handles all crime within that city's jurisdiction. Consequently, we used this figure as the overall homicide rate of the city.

Quantitative Predictor Variables

1. Unemployment rate: Unemployment rate is the percentage of unemployed residents within a city's labor force. The United States Bureau of Labor Statistics provides publically available datasets for the unemployment rate in metropolitan areas. In 2010, the United States overall unemployment rate was 9.6, but it varied from 3.9 percent to 28.8 percent across United States cities.
2. Population density: We found population density by dividing the city's total population by its total area (in kilometers squared). The values for both population and area were from the US Census Bureau's 2010 census.
3. Average Temperature: We used the mean temperature in degrees Fahrenheit during 2010 as recorded by the National Oceanic and Atmospheric Administration (NOAA). The NOAA records temperatures for major cities as well as for specific geographic areas of each state. For cities not in the NOAA online database, we approximated to the best of our ability by finding the mean temperature of the appropriate geographic region of the state.
4. Precipitation: Precipitation per city was recorded in inches by the NOAA. For cities not in the NOAA online database, we approximated to the best of our ability by finding the precipitation of the appropriate geographic region of the state.
5. Rate of Gun Ownership in State: Because we were unable to find consistent, reliable statistics on the rate of gun ownership in each city, we used the age adjusted rate of suicide firearm deaths per 100,000 as a proxy. This statistic has been documented as a reliable proxy for gun ownership and is available on a state level through the Center of Disease Control's Fatal Injury Reports.
6. Median Household Income: For median household income, we used the estimate from the US Census Bureau's 2010 American Community Survey. Median household income is based on the total number of households and families and includes those that have no income. The unit is 2010 inflation-adjusted dollars.
7. Percent of Population Identifying as Non-White: The 2010 US Census includes racial proportions of each city's population. To find the percentage of non-white residents, we subtracted the percentage of white residents from 100 percent.

8. Cost of Living: Although there are a few resources documenting the cost of living in various cities, we had trouble finding a credible, consistent source. As a result, we used gross median rent of the city as a proxy, since housing costs account for a large portion of the cost of living. Gross rent is expressed in 2010 dollars and includes the contract rent as well as an estimate for the monthly cost of utilities. The source was the US Census Bureau's 2010 American Community Survey.
9. Percent of Population with High School Diploma or Higher: Here, we found the percent of the population with an educational attainment level of at least a high school diploma from the US Census Bureau's 2010 American Community Survey.

Qualitative Predictive Variable

10. Census Region: The US Census Bureau designates four main regions of the United States: Northeast (NE), Midwest (MW), South (S), and West (W). The regions are defined as follows:
    a. Northeast (NE): Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, Pennsylvania
    b. Midwest (MW): Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
    c. South (S): Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee
    d. West (W): Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington

    We modeled this predictor with dummy variables coded (0, 1) with MW as the baseline level.

IV. Analysis

Note: All outputs and figures referenced in the text can be found in the Appendix

IV.A. Variable Plots

Before fitting a regression model, we first observed the distributions of our independent and dependent variables, as seen in **Supplemental Figures 1 and 2**. We found that the response variable, homicide rate, appears to be skewed right. A few predictor variables, specifically median gross rent and population density, appear to be skewed right as well. The predictor variable high school diploma attainment rate is the only variable that is skewed left. The rest seemed close to a normal distribution.

IV.B. Correlation Analysis

We proceeded by examining the correlation coefficients between each quantitative independent variable and homicide rate, displayed in **Supplemental Figure 4**, to assess the relationships between the variables. Based on the scatter plots, the percentage of population identifying as non-white appears to exhibit fairly strong correlation with homicide rate ($R = 0.71$), whereas other variables, such as unemployment rate and percent of population with high school diploma, seem to have very little obvious trend ($R < 0.25$). There were several predictor variables that had negative correlation (Gun Ownership, Median Gross Rent, and HS Diploma Attainment Rate) and several with positive correlation (Population Density, Average Temperature, Precipitation, Unemployment Rate, Median Household Income, and Percent Non-white). While many of these relationships might seem intuitively correct, such as HS Diploma Attainment being negatively correlated with Homicide Rate, the magnitudes are not large and none can indicate causality.

IV.C. Model Selection and Screening Techniques

We began the model screening process with all ten variables mentioned in the data description by running simple heuristic methods, particularly both-ways stepwise regression, shown in **Output 3,** and best subsets regression, shown in **Outputs 1 and 2**. In the initial iteration, we included no interaction terms and did not perform any transformations on the data. For stepwise regression, the alpha value required to enter and to remove were 0.05 and 0.10, respectively.

We decided to move forward with both-ways stepwise regression with interaction terms to see whether or not this would improve the fit of the model. We used the same alpha to enter and alpha to stay as in the case without interaction variables. Using this method, we found the following model displayed in **Output 4**. We recognized that one of the interaction variables, Precipitation*Percent Non-white had an insignificant p-value of 0.197 and decided to rerun the model without this variable. This resulted in a model, **Output 5**, which we were fairly happy with at the end of these variable screening techniques. The multiple linear regression model had an R^2 value of 63.57% and adjusted R^2 of 61.22%. We kept non-significant variables in this model due to the Principle of Hierarchy. For example, Gun Ownership in State has a p-value of 0.391, which is far greater than both alpha to enter and alpha to stay. However, since Percent Non-white * Gun Ownership was a significant variable ($p < 0.10$), we needed to keep the Gun Ownership independent variable in our model.

We proceeded by running best subsets regression to generate more candidate models, as shown in **Output 6**. In this iteration of best subsets, we included all independent variables as well as several interaction variables that we chose intuitively. For example, we included Median Gross Rent * Median Household Income because rent and income are likely related and may be an appropriate interaction term. From the best subsets results, we selected a candidate model with a relatively high adjusted R^2, low Mallows Cp, and low standard deviation, and ran a multiple linear regression model, displayed in **Output 7**. We noticed that the model did not follow the principle of hierarchy. As a result, we reran the model, shown in **Output 8**, with the additional variables Median Gross Rent and Gun Ownership to maintain hierarchy. We also noticed that several of the predictors in **Output 8** had p values greater than 0.05, which suggests that many predictors are not significant.

After another glance at outputs including interaction terms, we noticed some things that stood out. The first and most obvious is the blatant multicollinearity. This, along with residual analysis and assumption-checking will be discussed in the next few sections.

IV.E. Multicollinearity

We noticed that adding in interaction terms may seem to increase model fit, but ultimately requires us to add more variables in to maintain hierarchy. This increases the multicollinearity of the entire model. For example, in the stepwise regression with second order interaction variables **(Output 4)**, the variance inflation factor (VIF) reached 94.55 for Percent Non-white, which is far too large. In fact, for all variables besides Median Gross Rent and Precipitation, the VIF was > 10. Though this model had the highest R^2 value we have seen, it should be noted that multicollinearity can inflate R^2 as well, so the increase in R^2 may be misleading. This tradeoff between wanting both a good fit and low multicollinearity led us to decide to avoid interaction terms altogether in order to minimize multicollinearity and have a parsimonious model. Though the R^2 value is not as high as when interaction terms are included, we believed that excluding interaction terms would lead us to the better model.

We examined our candidate models in **Outputs 2 and 3**, which had no interaction terms. We ultimately decided to proceed with the model in **Output 3**. Although the model in **Output 2** has a marginally higher adjusted R^2 of 57.17%, Precipitation has a p-value of 0.132, which is above our threshold of 0.05, suggesting that precipitation is not significant. Though the model in **Output 3** has a slightly lower adjusted R^2 of 56.59%, all of its predictors are significant, and we found it to be the more parsimonious model. It should also be noted that all predictors in this model have a VIF of about one, which is ideal.

Finally, looking at the correlation matrix of all predictor variables in **Output 11** reveals that no pairwise correlations have a R value near 0.90, a good sign of little to no multicollinearity in our basic model.

IV.F. Residual Analysis and Transformations

We then created residual graphs, displayed in **Figure 3**, for the model in **Output 3**. We noticed several problems from the residual plots. First, the qqplot and Anderson-Darling statistic of 1.600 (with a p-value <0.005) suggest the errors are not normally distributed. In particular, the tails stray from the normal distribution. The histogram confirms this, as there is a clear outlier among the residuals. Furthermore, the plot of residuals against fitted values violates the constant variance assumption, since there appears to be increasing error variance. These observations suggest that a transformation of the y variable may be appropriate to normalize the residuals and meet the assumptions.

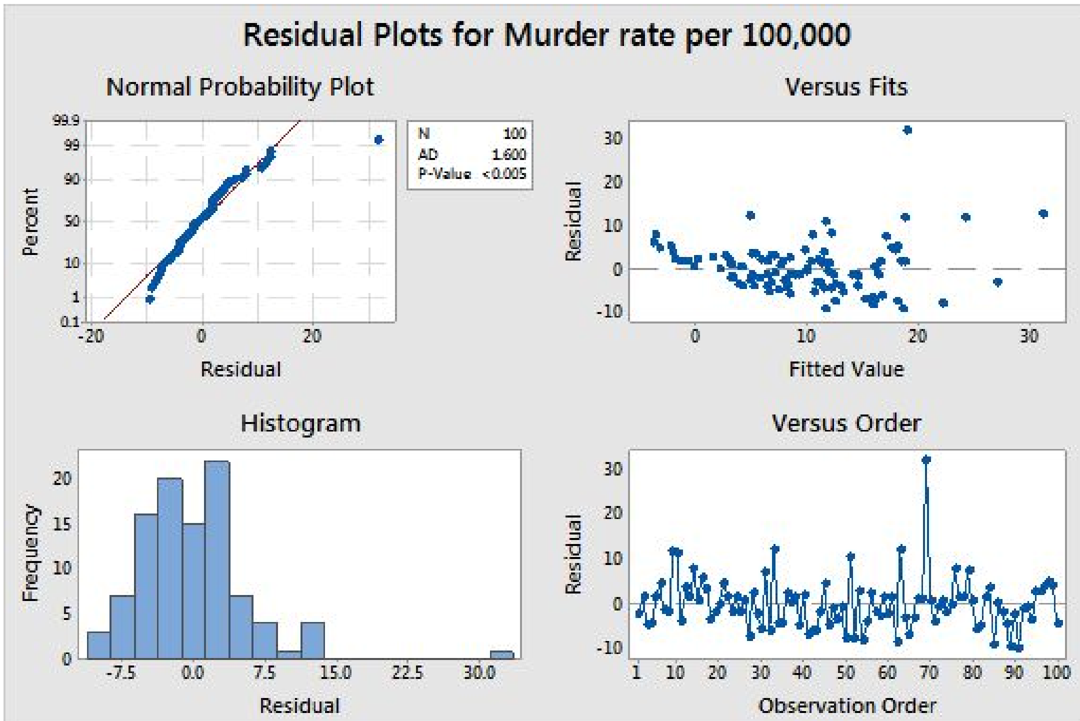


Figure 3: Initial Residual Plots

Thus, we tried two transformations: natural log and square root. It should be noted that there are a few cities with a homicide rate of 0, and natural log excludes those data points. After rerunning a model with the same screening methods with a transformed response variable, we arrived at two models with the same set of predictors as previously chosen. Upon examining their residual plots in **Supplemental Figures 8 and 9**, we find the previous issues have been adequately resolved—the qqplot, histogram, and AD statistic with p-values of 0.226 and 0.212, respectively, indicate normality. In all residual plots, including versus fits, the random clouds indicate both transformed models have correct functional form and identical distribution. To perform a two-tailed Durbin-Watson *d*-test, we have H_0: no residual correlation, H_a: positive or negative residual correlation. We compare *d* and (4 - *d*) against the lower and upper tabulated values of the rejection region; in this case for k = 3, n = 100, $d_{L,\,0.05}$ = 1.613 and $d_{U,\,0.05}$ = 1.736. The test statistic *d* for the LN model was 1.77394 > $d_{U,\,0.05}$ = 1.736, or in the non-rejection region. The test statistic *d* for the SQRT model was 2.00720 > $d_{U,\,0.05}$ = 1.736, also in the non-rejection region. While both models adequately pass all regression assumptions in residual analysis, **Outputs 9 and 10** show that the SQRT model has a slightly superior adjusted R^2 = 63.73% but also a higher Standard Error = 0.8397, while the LN model only had adjusted R^2 = 60.26% but a lower S = 0.5951. Ultimately, both regressions have F-values well above 40, corresponding to p-values ≈ 0, and are both equally valid models. We choose to continue working with the SQRT transformed model from here on.

IV.G. Assumption Checks

After transforming our response variable, our model meets all assumptions.

1. Expected value of residuals is zero

   Check: By inspection of the histogram and versus fits plot in **Supplemental Figure 9**, we see the mean of the error terms is zero.
2. Residuals are normally distributed

   Check: Based on the residual qqplot, histogram, and Anderson-Darling statistic, the residuals are fairly normal.
3. The errors are independent.

   Check: The plot of residuals against observation order as well as the Durbin-Watson statistic suggest independence.
4. The residuals have constant variance.

   Check: The scatter plot of residuals versus fitted values appears fairly random and suggests constant variance.

Also, since we have transformed our response variable, we would like to recheck multicollinearity once again. In our transformed model **(Output 10)**, VIF's for all predictors are close to one. Furthermore, the simple correlation coefficients between the predictors in the model are all less than

0.5, which suggests no significant correlation (**Output 11**). Therefore, we pass the multicollinearity test.

IV.H. Unusual Observations

We used Minitab's unusual observations diagnostics to identify six potential outliers and influential points, as shown in **Output 12**. We then assessed the leverage values, studentized residuals (SRES), studentized deleted residuals (TRES), and DFITS values for those points.

Because observations 4, 11, 63, and 69, have absolute SRES and absolute TRES values greater than two, they are most likely outliers. All observations have Cook's Distance Values of less than one and DFITS values less than two, suggesting that none of them are influential. However, observations 59 and 63 do have relatively large leverage values, and are thus, high leverage points. To correct our model for the unusual observations, we began deleting outliers one by one. It should be noted that the two cities with a homicide rate of zero were both outliers. This makes sense, as it is quite rare for cities to have no homicides within a year. After deleting one outlier, we noticed that the p-value for the Anderson Darling test slightly decreased, suggesting that the deletion decreased the normality of the residuals. Further deletions were consistent with this point. As a result, we decided to delete only one outlier with the greatest SRES and TRES, improving our model's adjusted R^2 increased from 63.73 percent in **Output 10** to 64.67 percent in **Output 13**, while still maintaining normality.

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.790000	65.75%	64.67%	64.5694	62.70%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	3.196	0.542	(2.120, 4.271)	5.90	0.000	
Gun Ownership in State	-0.0903	0.0325	(-0.1548, -0.0258)	-2.78	0.007	1.23
Percent Non-white	0.06062	0.00476	(0.05117, 0.07008)	12.73	0.000	1.02
Median Gross Rent (CoL)	-0.002579	0.000450	(-0.003472, -0.001685)	-5.73	0.000	1.25

Regression Equation

SQRT(Murder rate) = 3.196 - 0.0903 Gun Ownership in State + 0.06062 Percent Non-white - 0.002579 Median Gross Rent (CoL)

Figure 4: Transformed Linear Regression Model with Unusual Observation Removed

V. Conclusions and Recommendations

To assess our final model, we found 2010 data for ten random cities not in our original dataset and used our model to predict the homicide rate in each city. After finding the predicted homicide rate for each city, we calculated the difference in percentage from the actual value of homicide rate for that city. On average, we were 7.965% off the actual value for homicide rate. For 5 of the 10 cases, the actual value for homicide rate was within our model's 90 percent prediction intervals. We believe this disparity in actual versus predicted homicide rate may be attributed to insufficient sample size. For future studies, we recommend collecting data on more cities.

We are cautious in making our recommendations about homicide rate because we are aware that correlation does not mean causation. We think that this analysis can serve as the first steps to an experimental design where cities test how changes in an independent variable affect homicide rate. Some predictor variables that we used are inherent to the city, such as cost of living, and trying to artificially manipulate the cost of living would imply gentrification of cities. Instead, a potential avenue would be to manipulate Gun Ownership with stronger gun regulations to see what the effect on the homicide rate in a city can be.

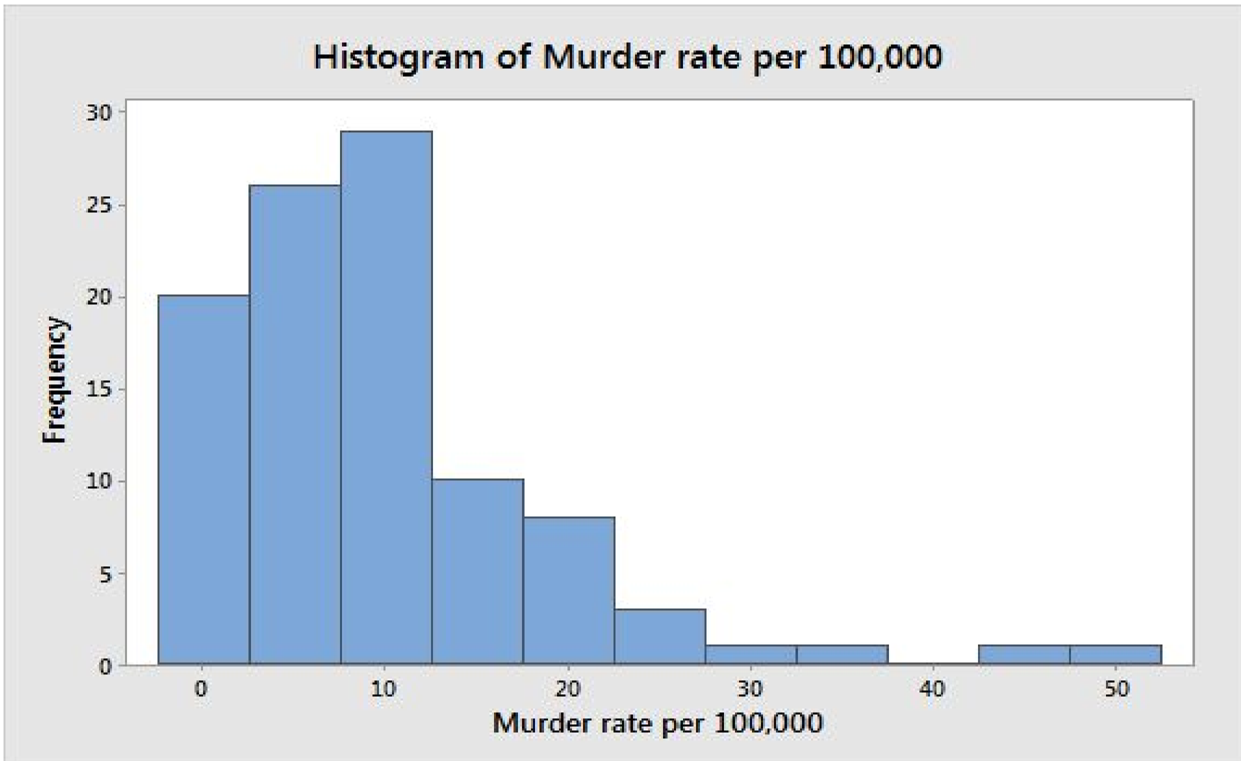
Improvements to data collection on other environmental variables could provide future studies such as this with more accurate and varied predictions. Furthermore, the data used for this model was from seven years ago, which means this data is slightly dated. We were limited by using census years, and rerunning the models on 2020 data would be a prudent next step if we want to use these factors to predict homicide rate.

VI. Appendix

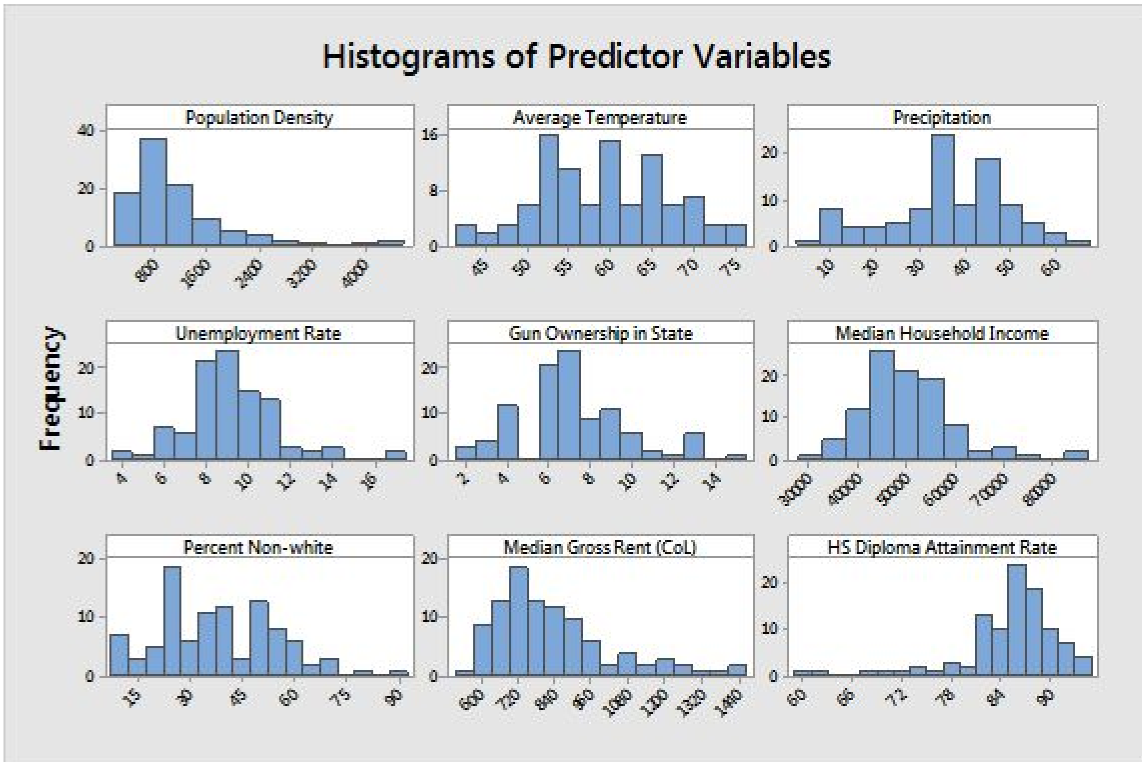
VI.A. Tools

Software: Microsoft Excel, Citrix, Minitab, R

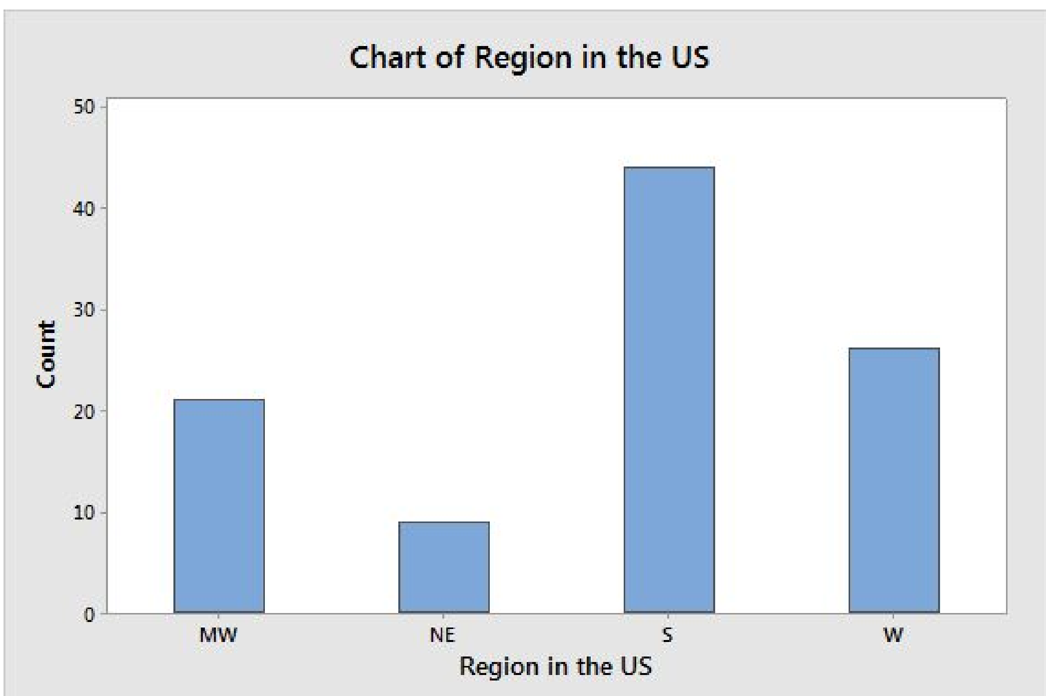
VI.B. Supplemental Figures



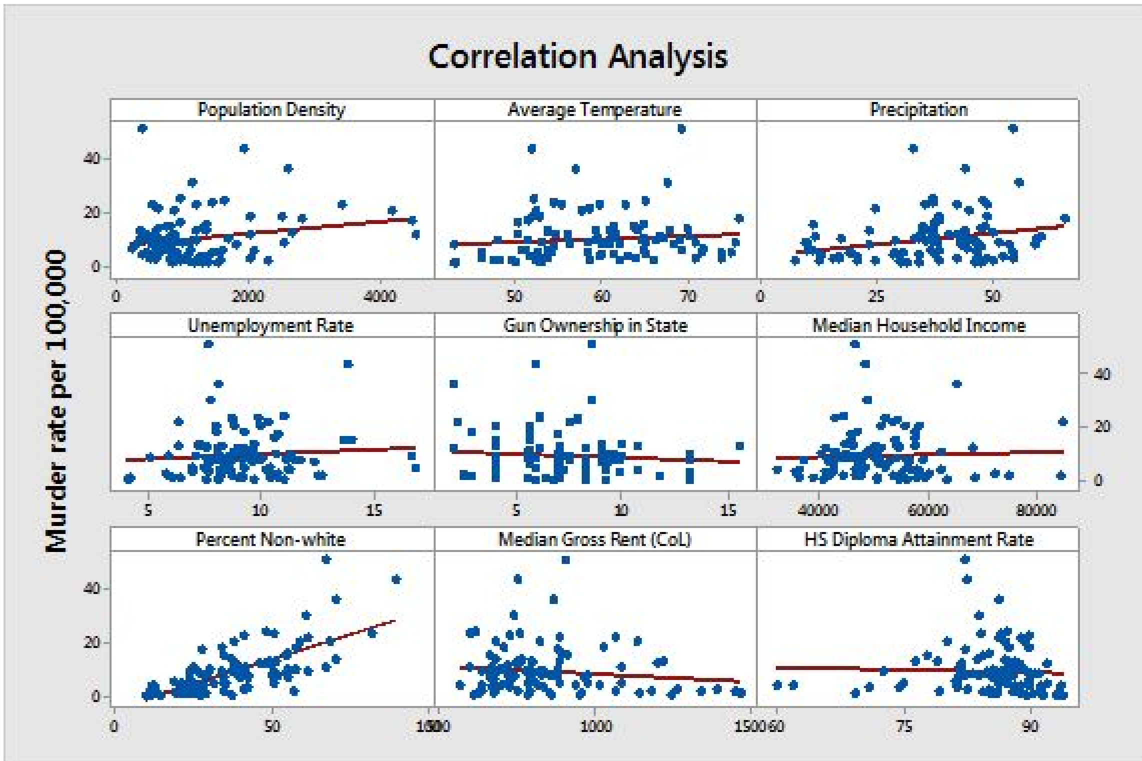
Supplemental Figure 1



Supplemental Figure 2



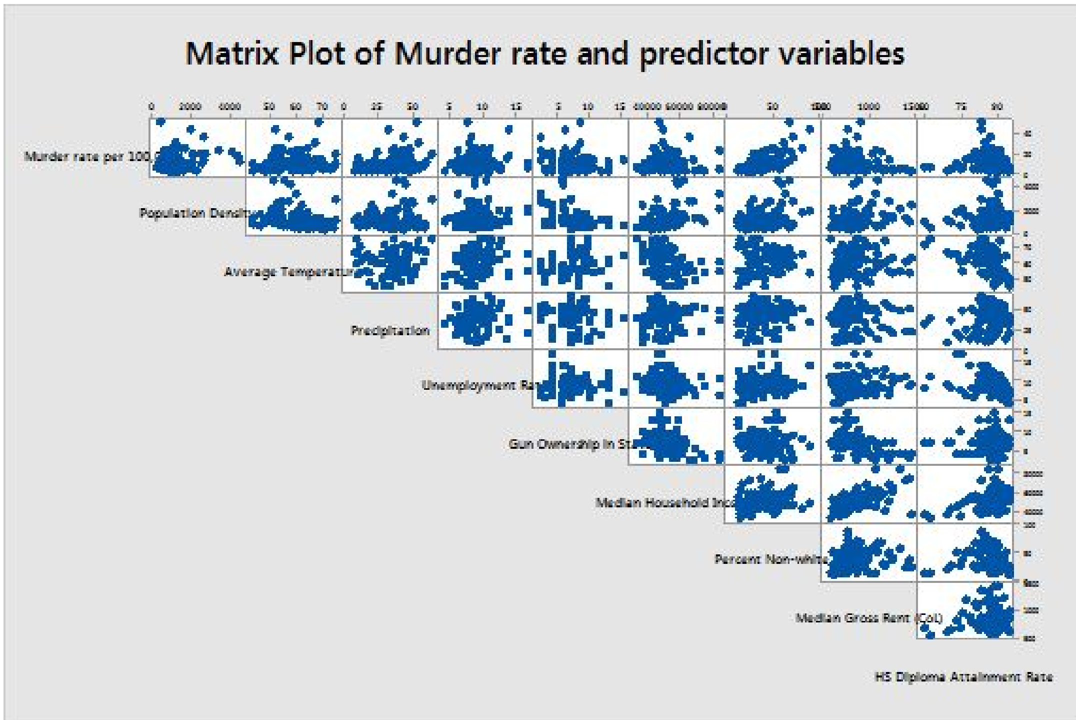
Supplemental Figure 3



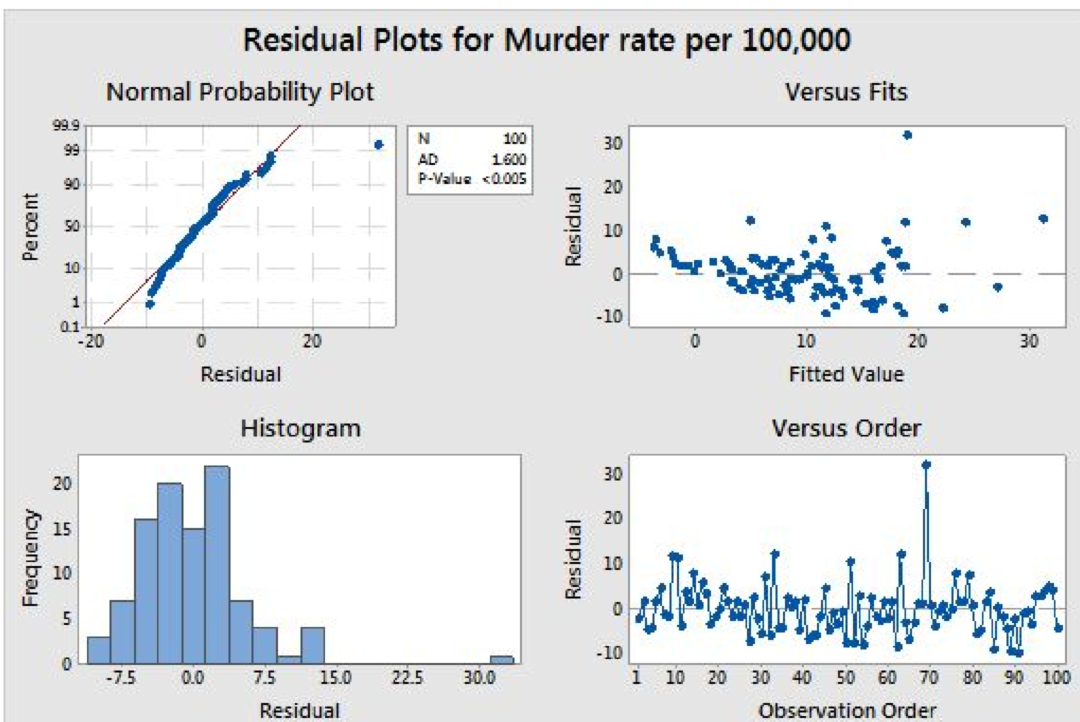
Supplemental Figure 4: Correlation Analysis Graph

Quantitative Predictor Variable	Correlation with Response Variable
Population Density	0.2238716
Average Temperature	0.09879202
Precipitation	0.2489102
Unemployment Rate	0.08407706
Gun Ownership in State	-0.08811079
Median Household Income	0.03923064
Percent Non-white	0.7101314
Median Gross Rent	-0.1256291
HS Diploma Attainment Rate	-0.04455943

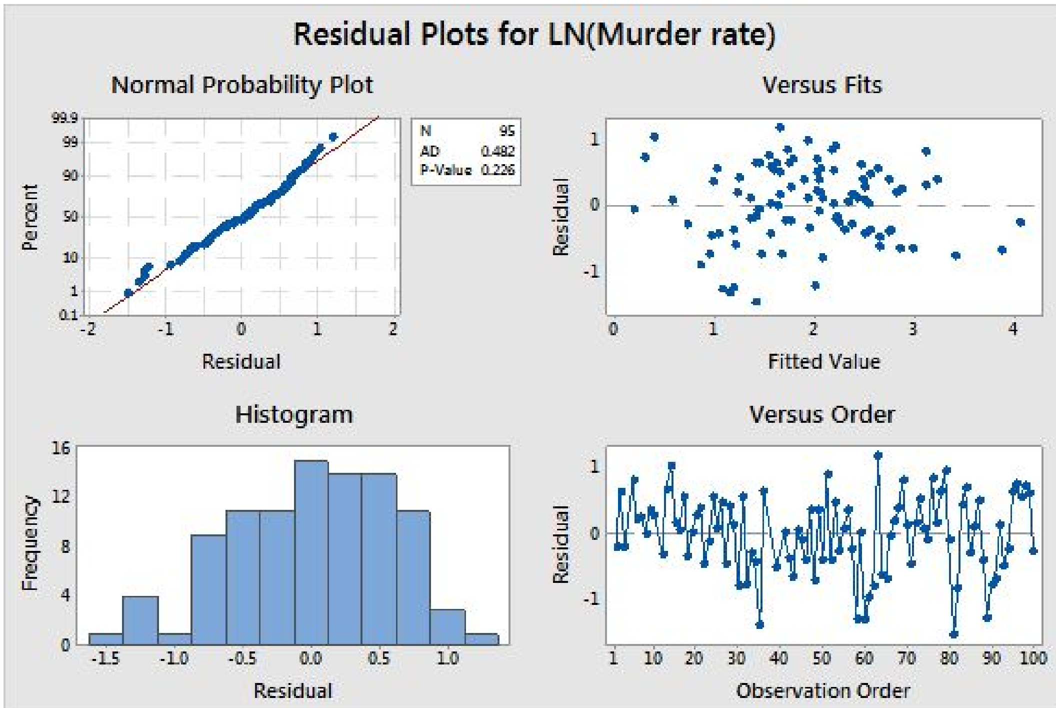
Supplemental Figure 5: Correlation Analysis Output



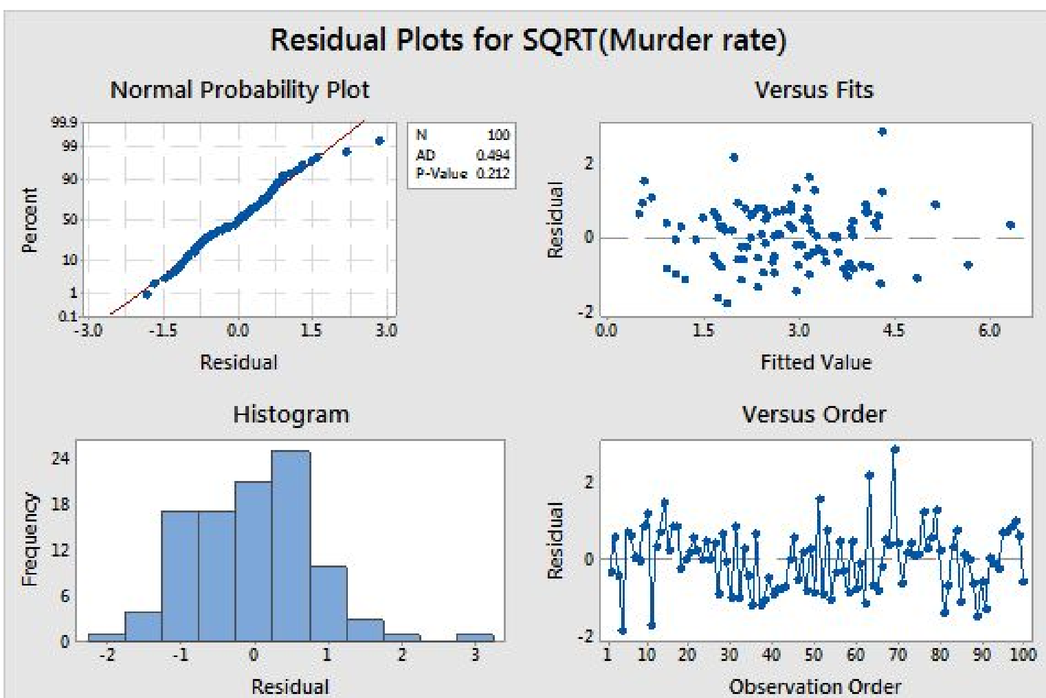
Supplemental Figure 6: Matrix Plot of All Variables



Supplemental Figure 7: Initial Residual Analysis Plots



Supplemental Figure 8: Transformed Residual Analysis Plots (Natural Log)



Supplemental Figure 9: Transformed Residual Analysis Plots (Square Root)

VI.C. Minitiab Outputs

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	s i t y	t u r e	t i o n	R a t e	t a t e	c o m e	h i t e	C o L)	R a t e
1	50.4	49.9	47.3	14.0	6.3181							X		
1	6.2	5.2	2.6	112.1	8.6913			X						
2	55.8	54.9	52.3	4.1	5.9959							X	X	
2	51.6	50.6	47.5	13.5	6.2777			X				X		
3	57.9	56.6	53.7	1.4	5.8827					X		X	X	
3	56.8	55.5	52.3	3.8	5.9570			X				X	X	
4	58.9	57.2	53.6	1.2	5.8432			X		X		X	X	
4	58.2	56.4	53.0	2.8	5.8936				X	X		X	X	
5	59.2	57.0	52.9	2.6	5.8537			X	X	X		X	X	
5	59.0	56.8	52.1	3.0	5.8691	X		X		X		X	X	
6	59.3	56.6	51.3	4.4	5.8798	X		X	X	X		X	X	
6	59.2	56.6	51.5	4.5	5.8814			X	X	X	X	X	X	
7	59.4	56.3	50.2	6.2	5.9041	X		X	X	X	X	X	X	
7	59.4	56.3	50.4	6.2	5.9045	X	X	X	X	X		X	X	
8	59.4	55.8	49.3	8.1	5.9337	X	X	X	X	X	X	X	X	
8	59.4	55.8	49.9	8.1	5.9348	X	X	X	X	X		X	X	X
9	59.4	55.4	48.9	10.0	5.9636	X	X	X	X	X	X	X	X	X

Output 1: Best Subsets Regression with no Interaction Terms

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5.84316	58.90%	57.17%	53.56%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.68	4.23	1.81	0.073	
Precipitation	0.0684	0.0451	1.52	0.132	1.04
Gun Ownership in State	-0.524	0.240	-2.19	0.031	1.23
Percent Non-white	0.3765	0.0355	10.61	0.000	1.07
Median Gross Rent (CoL)	-0.01360	0.00333	-4.09	0.000	1.25

Regression Equation

Murder rate per 100,000 = 7.68 + 0.0684 Precipitation - 0.524 Gun Ownership in State
+ 0.3765 Percent Non-white - 0.01360 Median Gross Rent (CoL)

Output 2: Candidate model from the result of Best Subsets (Output 1)

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.88273  57.90%     56.59%      53.68%


Coefficients

Term                           Coef  SE Coef  T-Value  P-Value   VIF
Constant                       9.86     4.01     2.46    0.016
Gun Ownership in State       -0.527    0.241    -2.18    0.031  1.23
Percent Non-white            0.3875   0.0350    11.08    0.000  1.02
Median Gross Rent (CoL)    -0.01373  0.00335    -4.10    0.000  1.24


Regression Equation

Murder rate per 100,000 = 9.86 - 0.527 Gun Ownership in State + 0.3875 Percent Non-white
                          - 0.01373 Median Gross Rent (CoL)
```

Output 3: Result of Stepwise Regression with no Interaction Terms

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.45055  65.74%     62.73%      55.36%


Coefficients

Term                                          Coef   SE Coef  T-Value  P-Value    VIF
Constant                                      -8.6      10.8    -0.79    0.430
Precipitation                               -0.044     0.111    -0.40    0.691   7.24
Gun Ownership in State                       0.775     0.642     1.21    0.230  10.11
Percent Non-white                            0.679     0.312     2.18    0.032  94.55
Median Gross Rent (CoL)                    0.00506   0.00792     0.64    0.525   8.11
Percent Non-white*Percent Non-white        0.00281   0.00160     1.75    0.083  18.91
Precipitation*Percent Non-white            0.00379   0.00292     1.30    0.197  22.10
Gun Ownership in State*Percent Non-white   -0.0348    0.0145    -2.40    0.018  19.47
Percent Non-white*Median Gross Rent (CoL) -0.000526  0.000213    -2.47    0.016  44.22


Regression Equation

Murder rate per 100,000 = -8.6 - 0.044 Precipitation + 0.775 Gun Ownership in State
                          + 0.679 Percent Non-white + 0.00506 Median Gross Rent (CoL)
```

Output 4: Stepwise Regression with Second Order Interaction Variables

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.55978  63.57%     61.22%      57.41%


Coefficients

Term                                            Coef   SE Coef  T-Value  P-Value    VIF
Constant                                        -8.7      10.1    -0.86    0.392
Median Gross Rent (CoL)                      0.00535   0.00808     0.66    0.510   8.11
Percent Non-white                              0.760     0.305     2.49    0.014  87.06
Gun Ownership in State                         0.548     0.636     0.86    0.391   9.54
Percent Non-white*Percent Non-white          0.00327   0.00161     2.03    0.045  18.26
Median Gross Rent (CoL)*Percent Non-white  -0.000540  0.000218    -2.48    0.015  44.19
Percent Non-white*Gun Ownership in State     -0.0279    0.0144    -1.94    0.056  18.46


Regression Equation

Murder rate per 100,000 = -8.7 + 0.00535 Median Gross Rent (CoL) + 0.760 Percent Non-white
                          + 0.548 Gun Ownership in State
                          + 0.00327 Percent Non-white*Percent Non-white
                          - 0.000540 Median Gross Rent (CoL)*Percent Non-white
                          - 0.0279 Percent Non-white*Gun Ownership in State
```

Output 5: Both-ways Stepwise Regression with Interaction Variables, Cleaned Up

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	ensity	rature	tation	t Rate	State	Income	-white	(CoL)	t Rate	* (Me	ent Ra	ecipit	n Dens	t Non-
1	50.4	49.9	47.3	19.4	6.3181							X							
1	18.0	17.2	12.9	94.9	8.1256														X
2	55.8	54.9	52.3	8.9	5.9959							X	X						
2	54.7	53.7	50.8	11.6	6.0734							X			X				
3	58.8	57.5	54.4	4.0	5.8197							X	X						X
3	58.0	56.7	53.4	5.9	5.8784							X			X				X
4	60.7	59.0	55.0	1.6	5.7171							X	X				X		X
4	60.2	58.5	54.8	2.7	5.7509			X				X	X						X
5	61.6	59.5	55.3	1.5	5.6787						X	X			X		X		X
5	61.1	59.1	55.0	2.5	5.7127					X		X	X				X		X
6	62.2	59.7	55.1	2.1	5.6671		X	X				X	X				X		X
6	62.0	59.6	55.1	2.5	5.6780					X	X	X			X		X		X
7	62.9	60.0	55.2	2.5	5.6447		X	X			X	X			X		X		X
7	62.5	59.7	54.5	3.3	5.6691		X	X	X			X	X				X		X
8	63.1	59.9	54.1	3.9	5.6542		X	X			X	X			X	X	X		X
8	63.1	59.9	54.2	3.9	5.6555		X	X	X		X	X			X		X		X
9	63.3	59.6	53.4	5.6	5.6755		X	X		X	X	X			X	X	X		X
9	63.2	59.6	53.5	5.6	5.6770		X	X	X	X	X	X			X		X		X
10	63.3	59.2	53.1	7.4	5.7019		X	X	X		X	X		X	X	X	X		X
10	63.3	59.2	52.8	7.4	5.7029		X	X	X	X	X	X		X	X		X		X
11	63.5	58.9	52.5	9.1	5.7248		X	X	X	X	X	X		X	X	X	X		X
11	63.4	58.8	52.1	9.3	5.7320		X	X	X		X	X	X	X	X	X	X		X
12	63.5	58.5	51.3	11.0	5.7546		X	X	X	X	X	X	X	X	X	X	X		X
12	63.5	58.4	51.1	11.1	5.7574	X	X	X	X	X	X	X		X	X	X	X		X
13	63.5	58.0	49.8	13.0	5.7872	X	X	X	X	X	X	X	X	X	X	X	X		X
13	63.5	58.0	51.1	13.0	5.7877		X	X	X	X	X	X	X	X	X	X	X	X	X
14	63.5	57.5	49.0	15.0	5.8212	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Output 6: Best Subsets Regression with Interaction Terms

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.64472  62.85%     60.03%      55.17%


Coefficients

Term                                   Coef   SE Coef  T-Value  P-Value    VIF
Constant                               10.9      14.1     0.77    0.442
Average Temperature                  -0.349     0.212    -1.64    0.104   8.50
Precipitation                        -0.598     0.339    -1.77    0.081  63.17
Median Household Income            0.000310  0.000143     2.17    0.033   5.59
Percent Non-white                    0.4902    0.0508     9.64    0.000   2.35
(Median Household Income) * (Me   -0.000000  0.000000    -4.02    0.000   5.01
(Average Temperature * Precipit     0.01095   0.00542     2.02    0.046  74.12
(Gun Ownership) * (Percent Non-    -0.01649   0.00523    -3.15    0.002   2.37


Regression Equation

Murder rate per 100,000 = 10.9 - 0.349 Average Temperature - 0.598 Precipitation
                          + 0.000310 Median Household Income + 0.4902 Percent Non-white
                          - 0.000000 (Median Household Income) * (Me
                          + 0.01095 (Average Temperature * Precipit
                          - 0.01649 (Gun Ownership) * (Percent Non-
```

Output 7: Candidate Model from the Result of Best Subsets (Output 6)

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.69625  63.00%     59.30%      53.25%


Coefficients

Term                                   Coef   SE Coef  T-Value  P-Value    VIF
Constant                                5.6      20.8     0.27    0.787
Average Temperature                  -0.321     0.221    -1.45    0.150   9.03
Precipitation                        -0.538     0.359    -1.50    0.138  69.77
Median Household Income            0.000310  0.000258     1.21    0.231  17.78
Percent Non-white                     0.542     0.104     5.22    0.000   9.63
(Median Household Income) * (Me   -0.000000  0.000000    -1.29    0.201  51.53
(Average Temperature * Precipit     0.01009   0.00571     1.77    0.080  80.56
(Gun Ownership) * (Percent Non-     -0.0240    0.0140    -1.72    0.089  16.61
Median Gross Rent (CoL)              0.0012    0.0126     0.10    0.923  18.85
Gun Ownership in State                0.377     0.647     0.58    0.562   9.39


Regression Equation

Murder rate per 100,000 = 5.6 - 0.321 Average Temperature - 0.538 Precipitation
                          + 0.000310 Median Household Income + 0.542 Percent Non-white
                          - 0.000000 (Median Household Income) * (Me
                          + 0.01009 (Average Temperature * Precipit
                          - 0.0240 (Gun Ownership) * (Percent Non-
                          + 0.0012 Median Gross Rent (CoL) + 0.377 Gun Ownership in State
```

Output 8: Result of Best Subsets (Output 6), Hierarchical Terms Included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.595091	61.53%	60.26%	57.33%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.227	0.324	3.79	0.000	
Precipitation	0.00936	0.00464	2.02	0.047	1.05
Percent Non-white	0.04061	0.00378	10.75	0.000	1.07
Median Gross Rent (CoL)	-0.001484	0.000317	-4.68	0.000	1.02

Regression Equation

LN(Murder rate) = 1.227 + 0.00936 Precipitation + 0.04061 Percent Non-white
- 0.001484 Median Gross Rent (CoL)

Output 9: Both-ways Stepwise Regression on LN Transformation, No Interaction Vars

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.839714	64.83%	63.73%	61.69%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.991	0.573	5.22	0.000	
Gun Ownership in State	-0.0810	0.0344	-2.35	0.021	1.23
Percent Non-white	0.06355	0.00499	12.73	0.000	1.02
Median Gross Rent (CoL)	-0.002512	0.000478	-5.25	0.000	1.24

Regression Equation

SQRT(Murder rate) = 2.991 - 0.0810 Gun Ownership in State + 0.06355 Percent Non-white
- 0.002512 Median Gross Rent (CoL)

Output 10: Both-ways Stepwise Regression on SQRT Transformation, No Interaction Vars

	Murder rate per	Population Densi	Average Temperat	Precipitation
Population Densi	0.224			
	0.025			
Average Temperat	0.099	-0.229		
	0.328	0.022		
Precipitation	0.249	0.087	0.090	
	0.013	0.387	0.373	
Unemployment Rat	0.084	0.121	0.193	0.043
	0.406	0.231	0.055	0.670
Gun Ownership in	-0.088	-0.428	0.143	-0.014
	0.383	0.000	0.155	0.891
Median Household	0.039	0.447	-0.362	0.044
	0.698	0.000	0.000	0.666
Percent Non-whit	0.710	0.301	0.168	0.204
	0.000	0.002	0.094	0.042
Median Gross Ren	-0.126	0.309	0.071	0.008
	0.213	0.002	0.482	0.941
HS Diploma Attai	-0.045	0.118	-0.540	0.159
	0.660	0.241	0.000	0.113

	Unemployment Rat	Gun Ownership in	Median Household	Percent Non-whit	Median Gross Ren
Gun Ownership in	-0.123				
	0.222				
Median Household	-0.178	-0.384			
	0.077	0.000			
Percent Non-whit	0.211	-0.078	0.201		
	0.035	0.441	0.045		
Median Gross Ren	0.135	-0.428	0.567	0.146	
	0.181	0.000	0.000	0.146	
HS Diploma Attai	-0.301	0.039	0.389	-0.061	0.092
	0.002	0.701	0.000	0.549	0.363

Cell Contents: Pearson correlation
P-Value

Output 11: Correlation Matrix of All Variables

Fits and Diagnostics for Unusual Observations

Obs	SQRT(MURDER)	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
4	0.000	1.853	0.118	(1.620, 2.087)	-1.853	-2.23	-2.28	0.019638
11	0.000	1.702	0.208	(1.289, 2.115)	-1.702	-2.09	-2.13	0.061387
45	1.049	0.473	0.306	(-0.134, 1.080)	0.576	0.74	0.73	0.132449
59	3.592	3.121	0.293	(2.540, 3.703)	0.470	0.60	0.60	0.121724
63	4.123	1.977	0.103	(1.772, 2.182)	2.146	2.58	2.66	0.015132
69	7.134	4.300	0.176	(3.951, 4.650)	2.834	3.45	3.67	0.044061

Obs	Cook's D	DFITS	
4	0.02	-0.322248	R
11	0.07	-0.544843	R
45	0.02	0.286948	X
59	0.01	0.221694	X
63	0.03	0.329182	R
69	0.14	0.787723	R

R Large residual
X Unusual X

Output 12: Unusual Observations Diagnostics

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.790000	65.75%	64.67%	64.5694	62.70%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	3.196	0.542	(2.120, 4.271)	5.90	0.000	
Gun Ownership in State	-0.0903	0.0325	(-0.1548, -0.0258)	-2.78	0.007	1.23
Percent Non-white	0.06062	0.00476	(0.05117, 0.07008)	12.73	0.000	1.02
Median Gross Rent (CoL)	-0.002579	0.000450	(-0.003472, -0.001685)	-5.73	0.000	1.25

Regression Equation

SQRT(Murder rate) = 3.196 - 0.0903 Gun Ownership in State + 0.06062 Percent Non-white
- 0.002579 Median Gross Rent (CoL)

Output 13: Regression Model After Removing Outlier

VII. Sources

Murder Rate

https://ucr.fbi.gov/crime-in-the-u.s

Gun Ownership by State

https://webappa.cdc.gov/sasweb/ncipc/mortrate.html

HS Diploma Achievement Rate

https://www.census.gov/library/publications/2010/compendia/databooks/educ.html

Median Household Income:

https://www2.census.gov/programs-surveys/acs/summary_file/2010/data/1_year_entire_sf/

US Census Region

https://www2.census.gov/geo/docs/maps-data/maps/reg_div.txt

Average Temperature

https://www.ncdc.noaa.gov/cag/time-series/us

Average Precipitation

https://www.ncdc.noaa.gov/cag/time-series/us

Gross Median Rent

https://www.socialexplorer.com/data/ACS2010/metadata/?ds=SE&table=T104

Percent Non-White

https://factfinder.census.gov

Unemployment Rate

https://data.bls.gov/cgi-bin/dsrv

Population Density

https://www.census.gov