ORIE 3120 Final Report

Analysis of Community Violent Crime Rates Using Socioeconomic Features

Group 4: Kashmala Arif, Ashna Gupta, Nayana Venukanthan, Kody Yang

Table of Contents:

I. INTRODUCTION

The US Census Bureau has released a comprehensive dataset, sourced from the UCI database, combining socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR, providing detailed insights into the characteristics of communities within the United States. With 1,994 instances each representing a community and 127 features ***(refer to Appendix A for description of variables)***, this dataset presents an invaluable opportunity to delve into the socio-economic landscape and crime dynamics across different communities.

A distinctive feature of this dataset is the normalization of all numeric data into the decimal range 0.00-1.00, facilitating uniformity and comparability across variables without altering their distribution and skew. This characteristic underscores the complexity of interpreting relationships between different attributes while simplifying within-attribute comparisons.

We aim to leverage this dataset to develop predictive models for crime rates within these communities. By analyzing the interplay between socioeconomic factors and crime data, we aim to answer the following questions:

1. What is the nature of the relationship between socioeconomic variables and violent crime rates?
2. What are the main features that determine the violent crime rate in a community?
3. Can we predict whether a community is safe or not using socioeconomic variables?

Policymakers can utilize the data-driven insights generated to formulate evidence-based policies addressing socio-economic disparities and reducing crime rates within communities. Law enforcement agencies can leverage the predictive models to allocate resources effectively and prioritize areas with higher forecasted crime rates for targeted interventions. Business owners can use the predictive models to gauge areas of low crime to conduct business and ensure their customers' safety.

II. EXPLORATORY DATA ANALYSIS

When ranking the top 15 correlation values for features with the violent crime rates per population, features `PctKids2Par`, `PctIlleg`, `PctFam2Par`, `racePctWhite`, and `PctYoungKids2Par` show the highest correlations. Notably, `PctKids2Par`, `PctIlleg`, `PctFam2Par`, and `PctTeen2Par` exhibit significant intercorrelation (See Table 1).

Table 1: Top 15 Correlated Factors with ViolentCrimes

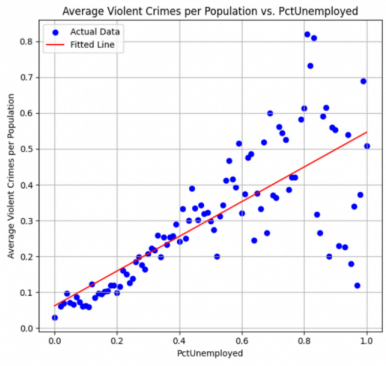
Feature Name	Feature Description	Correlation	P-Value
`ViolentCrimesPerPop`	total number of violent crimes per 100K population	1.00	N/A
`PctKids2Par`	% of kids in family housing with two parents	-0.738424	0.000
`PctIlleg`	% of population that is undocumented immigrants	0.737957	0.000
`PctFam2Par`	% of families (with kids) that are headed by two parents	-0.706667	0.000

racePctWhite	% of population that is caucasian	-0.684770	0.000
PctYoungKids2Par	% of kids 4 and under in two parent households	-0.666059	0.000
PctTeen2Par	% of kids age 12-17 in two parent households	-0.661582	0.000
racepctblack	% of population that is African American	0.631264	0.000
pctWInvInc	% of households with investment / rent income in 1989	-0.576324	0.000
pctWPubAsst	% of households with public assistance income in 1989	0.574665	0.000
FemalePctDiv	% of females who are divorced	0.556032	0.000
TotalPctDiv	% of population who are divorced	0.552777	0.000
PctPersOwnOccup	% of people in owner occupied households	-0.525491	0.000
MalePctDivorce	% of males who are divorced	0.525407	0.000
PctPopUnderPov	% of people under the poverty level	0.521877	0.000
PctUnemployed	% of people 16 and over, in the labor force, and unemployed	0.504235	0.000

III. LINEAR REGRESSION

At the start of our analysis, we explored linear regression as a foundational step to explore the predictions of community violent crime rates. When creating the scatter plots for the top 15 most highly correlated features to violent crime, we noticed a significant amount of noise and outliers, which hindered our ability to discern the nature of the relationship ***(Appendix B).*** Some of these outliers were infeasible, such as a 100% poverty rate, and were likely added to the dataset as a result of normalization.

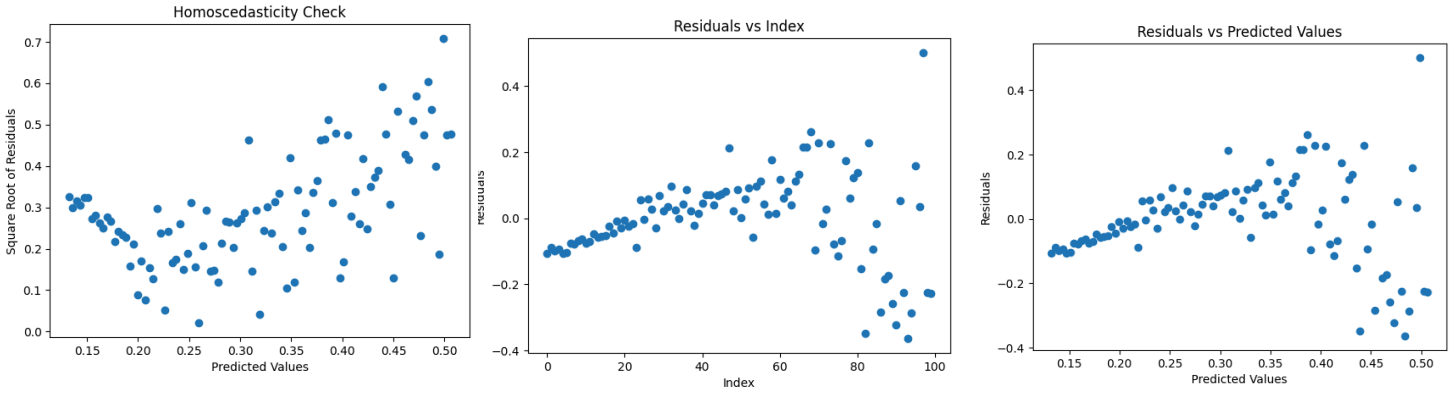
Therefore, we proceeded to plot these features against the average crime rates per population ***(Appendix C)***. Unsurprisingly, many of the correlations increased significantly, though some decreased. Interestingly, the charts showing decreased correlations indicated a polynomial trend rather than a linear one, suggesting more complex relationships in the data. The features with polynomial trends include **racepctblack**, **PctPopUnderPov**, and **PctUnemployed** where the relationships start off positively linear but inverse after around 60%.



It is important to note the ambiguity of causality here. While **racepctblack**, **PctPopUnderPov**, **PctUnemployed,** and **ViolentCrimesPerPop** are correlated, this does not necessarily imply causation. Other confounding variables could influence this relationship. For example, committing a crime can make it harder to find employment, leading to economic disadvantage. On the flip-side, those who are economically disadvantaged may be more compelled to engage in crimes, such as theft, as a means of survival.

TESTING LINEAR REGRESSION ASSUMPTIONS

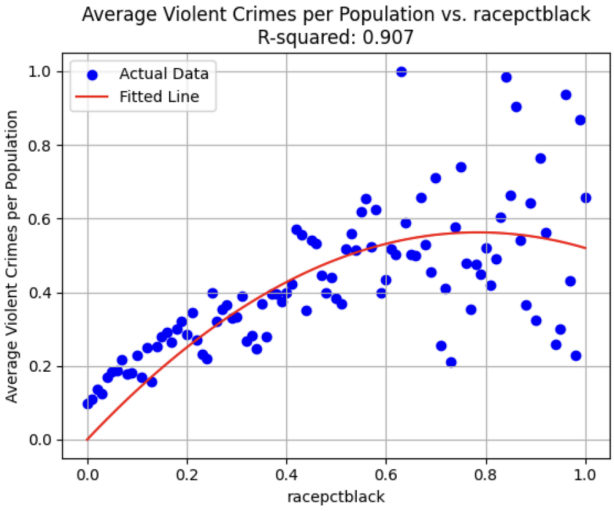
When we examined the residuals ***(Appendix D)*** for **racepctblack**, **PctPopUnderPov**, and **PctUnemployed**, it became evident that the assumptions of linear regression were violated. As shown below, the plot of residuals versus predicted values of **PctPopUnderPov** moves away from 0 as the predicted value of the **PctPopUnderPov** increases, indicating a violation of the assumption of homoscedasticity. Homoscedasticity (constant variance of residuals) is an important assumption of linear regression. Figure 1, depicting the plots for **PctPopUnderPov,** suggests that the variability of the residuals (the differences between the actual and predicted values) is not constant across all levels of the predicted values of **PctPopUnderPov**. This means that the model's errors (residuals) tend to increase or decrease systematically as the poverty percentage increases, indicating that the model might not be capturing some underlying pattern or relationship in the data. Practically, this could mean that the linear regression model is not the best model to use for predicting the average violent crimes per population based solely on **PctPopUnderPov**. It might be necessary to explore other models or consider additional variables that could improve the model's performance and address the issue of heteroscedasticity for all three of these features. In the Residuals vs Index graph for **PctPopUnderPov**, we can also see that the residuals vs index helps us assess whether the residuals are independent of each other. Since there is a pattern at the beginning of the graph, the assumption of independence is violated, and there are biased estimates leading to incorrect inference. So there are other factors that play a role in why a lower poverty rate is associated with lower crimes. In contrast, when the population poverty is at least 60%, there is no pattern, suggesting there are not as many biased estimates contributing to its prediction. However, it was particularly interesting that the trend changes after 60%, so we need to explore what factors cause this across the three features: **racepctblack,PctUnemployed,** and **PctPopUnderPov**.



Due to the violation in the assumptions of linear regression, we plan to further investigate these trends, with a focus on polynomial regression models. If these models demonstrate higher R^2 values compared to the other top 12 linear models, we will proceed with the polynomial approach.

IV. QUADRATIC REGRESSION

When graphing each of the 3 features: **`racepctblack`**, **`PctPopUnderPov`**, and **`PctUnemployed`** against average violent crimes per population ***(Appendix E)***, we found that **`racepctblack`** and **`PctPopUnderPov`** gave the highest R^2 values and very low mean squared error (MSE). These R^2 values were very high leading us to choose these models over other linear ones.



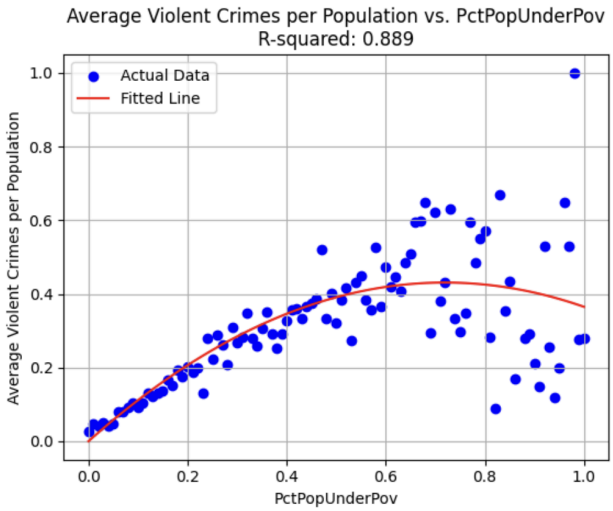


Table 2: Statistically Significant Features in Linear Regression for Violent Crime Rate

Feature	R^2	P-Value	MSE
racepctblack	0.907	0	0.0200603133830 9
PctPopUnderPov	0.889	0	0.0146248225638 2

While more sophisticated models, such as higher-degree polynomial regressions, could model the data with higher granularity, a quadratic model presents complexity while maintaining interpretability. Additionally, although higher-degree polynomial regressions would offer a closer fit to the data, they risk overfitting. This would make the model less generalizable and potentially less useful for predictive purposes across different communities.

In terms of applications, stakeholders such as policymakers and law enforcement can make use of these results for future predictions of violent crime rates based on demographics. However, from a business or general citizen perspective, rather than precise predictions of violent crime rate, it is more

valuable to know whether an area is relatively safe or unsafe to visit or conduct business in. Thus, we were curious to see whether top correlated features could be used in a binary logistic regression to predict whether an area has a high crime rate or not.

To further improve the model, we decided to explore multiple linear regression. We examined variables with scatter plots and R^2 values most closely aligned with the **racepctblack** and **PctPopUnderPov** plots. However, when exploring multiple linear regression, issues came about when we grouped the data to plot the average crime rates leading to mathematical errors ***(Appendix E)***. Additionally, the R^2 value for the multiple linear regression model was lower than the ones we currently have, so we won't be continuing with it for the purpose of this project.

V. LOGISTIC REGRESSION WITH TRAIN/TEST SPLIT

We sought to predict an area's safety based on this data using logistic regression. To develop a model, we first used sklearn's test_train_split() method to randomly split the data into portions by percentage: 80% for training and 20% for testing. We also needed to modify our target variable, **ViolentCrimesPerPop,** into a binary variable in order to make a Logistic Regression model implementable. The threshold for our new binary variable, **Unsafe_or_Not,** was a **ViolentCrimesPerPop** value of 0.50; in other words, a community is considered unsafe (**Unsafe_or_Not = 1)** when the violent crimes per population is above 0.50 (500 crimes per 1000 people) and safe (**Unsafe_or_Not = 0)** otherwise. We then used the training set to fit models beginning with the top 15 variables from Table 1 we found to be highly correlated with violent crime. We continuously fit logistic regression models until only statistically significant variables remained. After experimenting and adding quadratic variables, the variables we chose for the model that together were all statistically significant were **racepctblack**, **racepctblack_squared**, **pctIlleg**, **PctPopUnderPov**, and **PctPopUnderPov_squared**.

After fitting the model with our training data, the p-values and coefficients for the logistic model are displayed in Table 3. All the p-values were shown to be less than 0.05 with this combination, indicating that as a set of predictors, they are statistically significant in predicting **Unsafe_or_Not**.

Table 3: Results for Logistic Model:

Feature	P-Value	Pseudo R^2	Coefficient Value
const	0.000	0.3878	-6.3405
PctPopUnderPov	0.000		11.4591
racepctblack	0.003		3.0585
PctIlleg	0.000		4.5003
PctPopUnderPov_squared	0.000		-10.2577

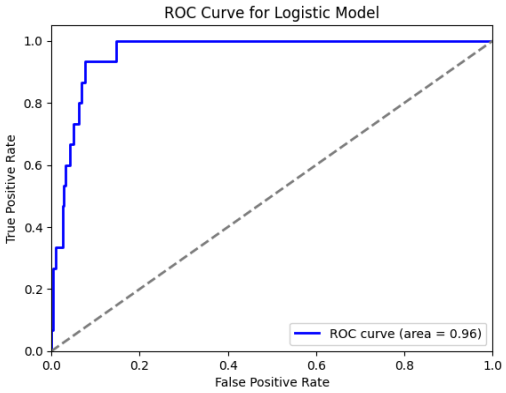
racepctblack_squared	0.028		-2.3383

To evaluate the prediction capabilities of our model, we then used the model to predict the 20% testing data that was remaining and extracted the performance metrics (see Table 4). Note that in this model, given the binary classification we created, a true positive is a correctly identified unsafe community, a true negative is a correctly identified safe community, a false positive is a safe community marked as unsafe, and a false negative is an unsafe community marked as safe.

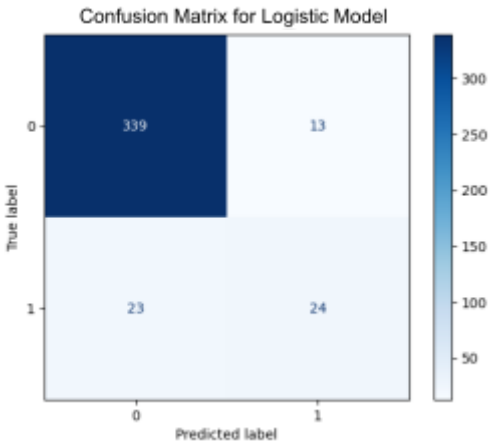
Table 4: Logistic Model Metrics

Metric	Value
Accuracy	0.967
Precision	0.750
Recall	0.200
F1 score	0.316
ROC AUC	0.962

Given that the model performed with a 96.7% accuracy, this implies that model performs well overall, unless there is some other factor such as class imbalance that is skewing the accuracy. The precision of the model is 75%, meaning 75% of all communities marked by the model as unsafe were actually unsafe. This is a relatively high precision, so the predictions are fairly reliable. The recall value is 20%, meaning that of all positive cases, only 20% were correctly identified as unsafe, and the remaining 80% were marked as safe (false negatives). This metric definitely raises questions about the performance of the model, as an overwhelming majority of unsafe communities were not correctly classified. The F1 score reinforces this, as the F1 score is the harmonic mean of precision and recall and is meant to provide a more holistic view of model performance. The value of 0.316 for a F1 score indicates that the model is not very reliable in correctly identifying unsafe communities.



In order to gain more insight into exactly how the model classified each binary class, we generated a receiving operating characteristic (ROC) curve, which is shown at right, which plots the diagnostic ability of a classifier as its threshold varies. Since our metrics show that the area under the ROC curve is 0.96, this is very close to 1 and suggests that there is a 96 % chance that the model correctly distinguishes between a random unsafe community and a random safe community.



Given the somewhat mixed results, we decided to delve further and generate a confusion matrix for our logistic model to gain insight into the model performance.

At left we see the confusion matrix for the model's prediction of the testing data. Here we can see a clear imbalance between the large number of safe communities and small number of unsafe communities, which explains the skewed accuracy and other metrics. More unsafe community data points are required to improve model performance, which is entirely random due to the

test train split. Typically, another solution would be adjusting the threshold, but we didn't see much difference in metrics when we tried this. For further investigation, the class imbalance should be addressed in order to improve model prediction.

VI. DISCUSSION & CONCLUSION

Question 1

Our analysis aimed to uncover the relationship between socioeconomic variables and violent crime rates. Initial exploration revealed a combination of both linear and non-linear patterns. Through regression analysis, we identified that the strongest model were quadratic trends between **racepctblack**, **PctPopUnderPov**, and average violent crimes per population. There are several explanations for these negative parabolic relationships, which peak at around 0.6 and decrease after. In areas with very low poverty rates, high levels of social cohesion and law enforcement can often mitigate crime. Additionally, a significant portion of violent crimes go unreported due to distrust in the police and the justice system (around 16% according to the National Crime Victimization Survey).

Question 2

Our quadratic regression results identified **racepctblack** and **PctPopUnderPov** as the main features that determine violent crime rates in a community. There are many reasons as to why these are the main features. **Racepctblack** has an R^2 value of 0.907. This suggests that approximately 90.7% of the variance in the violent crime rate can be explained by the percentage of the population that is African American. Similarly, with an R^2 value of 0.889, approximately 88.9% of the variance in the violent crime rate can be explained by **PctPopUnderPov**. The model with **racepctblack** has an MSE of 0.020. This indicates on average, the squared difference between the observed violent crime rates and the predicted values from the model is approximately 0.020. Similarly, for **PctPopUnderPov**, the MSE value is approximately 0.015. These low MSE values indicate that the regression models' predictions are extremely close to the actual violent crime rates.

Studies have shown "how laws and policies related to housing, land use, education, and transportation created enduring patterns of segregation in neighborhoods and schools" (Turner and Greene). Overall, due to the "Separate and Unequal" principle characterizing American neighborhoods, people of color (particularly black Americans and illegal immigrants) are overrepresented in high-poverty areas which are associated with the lack of essential resources which affects economic mobility and cyclically induces crime and bias in policing.

Question 3

Based on our findings, we can state socioeconomic variables, specifically **racepctblack**, **PctPopUnderPov**, and **PctIlleg** are influential in the binary classification of communities as safe or unsafe. Our analysis revealed that these three variables were highly predictive of community safety based on their p-values. However, the logistic model metrics suggest there is room for improvement in accurately identifying positive instances, in this case unsafe communities. As identifying unsafe communities is the main objective behind developing such a model, further analysis and development of this logistic model with more training data and creating more balanced classes should be done to better identify unsafe environments. Justifications for the correlation of **racepctblack** and

PctPopUnderPov with **ViolentCrimesPerPop** remain the same as observed above for Question 2. However, for the logistic regression model, the new feature of **PctIlleg** was identified as statistically significant as a predictor for the safety of communities.

This was particularly interesting, as while there is a general misconception in the U.S. that undocumented immigrants are likely to commit violent crimes, studies have proven that undocumented immigrants had substantially lower crime rates than native-born Americans and legal immigrants (Light). The correlation can possibly be attributed to the same economic and racial segregation of U.S. neighborhoods, which systematically prevents undocumented immigrants from settling in or moving to more affluent and safer neighborhoods (Turner and Greene). The inclusion of these features in the linear and logistic model may indicate potential unconscious racial and economic bias in machine learning models, which has been shown to be present in previous studies.

VII. CONCLUSION

We believe his study is significant in bringing to light the complex socio-economic dynamics characterized by American neighborhoods. While machine learning techniques alone cannot provide a holistic understanding of these dynamics, the models and results produced in this investigation reveal valuable insight into the relative correlations of various features and crime rates on the community-level. Understanding these correlations is crucial for policymakers, law enforcement, and community organizations as it emphasizes the need for comprehensive socio-economic interventions such as investing in education, fair housing initiatives, job creation and police reform to address the root causes of crime and reduce socio-economic and systemic disparities. By acknowledging and addressing these factors, stakeholders can work towards creating safer, more equitable communities.

For further investigation and to further enhance our analysis and explore more complex relationships, we can undertake other steps to refine our dataset. One important step we can take is removing outliers and cleaning the data to ensure its accuracy and reliability. This is because outliers can distort our analysis and lead to inaccurate conclusions. By removing them, we can get a more representative dataset that better reflects the underlying patterns and trends in the data. Additionally, we can conduct further investigation into the causality of the relationships between the socioeconomic variables and community safety. While our initial analysis identified strong predictors, understanding the causal mechanisms behind these relationships can provide valuable insights for policymakers. Furthermore, we can explore more advanced analysis methods such as multiple linear regression to examine the combined effect of multiple variables on community safety. This can help us identify the most significant factors and provide a better understanding of factors influencing community safety. In all, by refining our dataset, investigating causality, and employing more advanced analysis methods, we can gain a deeper understanding of the socioeconomic factors affecting community safety.

VIII. Works Cited

Light, Michael T., Jingying He, and Jason P. Robey. "Comparing Crime Rates Between Undocumented Immigrants, Legal Immigrants, and Native-born US Citizens in Texas." PNAS, vol. 117, no. 51, Dec. 2020, pp. 32340-32347. Office of Justice Programs, https://www.ojp.gov/library/publications/comparing-crime-rates-between-undocumented-immigrants-legal-immigrants-and.

Turner, Margery Austin and Solomon Greene. “Causes and Consequences of Separate and Unequal Neighborhoods.” *Urban.org*, n.d., https://www.urban.org/racial-equity-analytics-lab/structural-racism-explainer-collection/causes-and-consequences-separate-and-unequal-neighborhoods.

U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. "Violent Crime against the Elderly Reported by Law Enforcement in Michigan, 2005-2009." *Bureau of Justice Statistics*, Aug. 2010, https://bjs.ojp.gov/content/pub/pdf/vnrp0610.pdf.

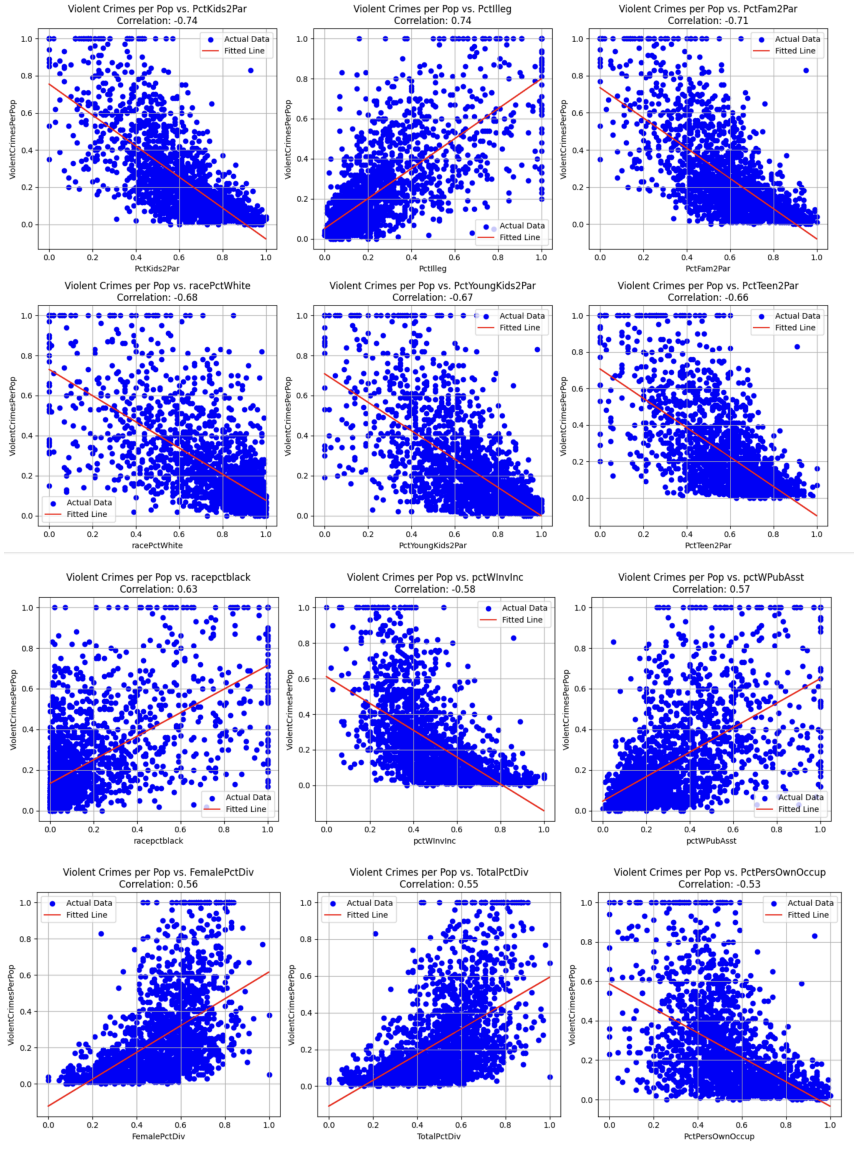
U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. "Victimizations Not Reported to the Police, 2006-2010." *Bureau of Justice Statistics*, Aug. 2012, https://bjs.ojp.gov/content/pub/pdf/vnrp0610.pdf

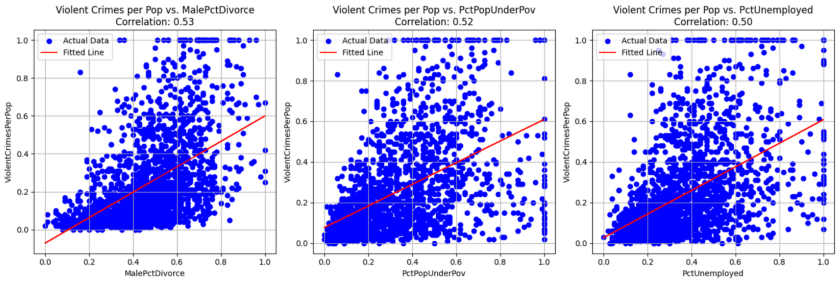
IX. APPENDIX

Appendix A

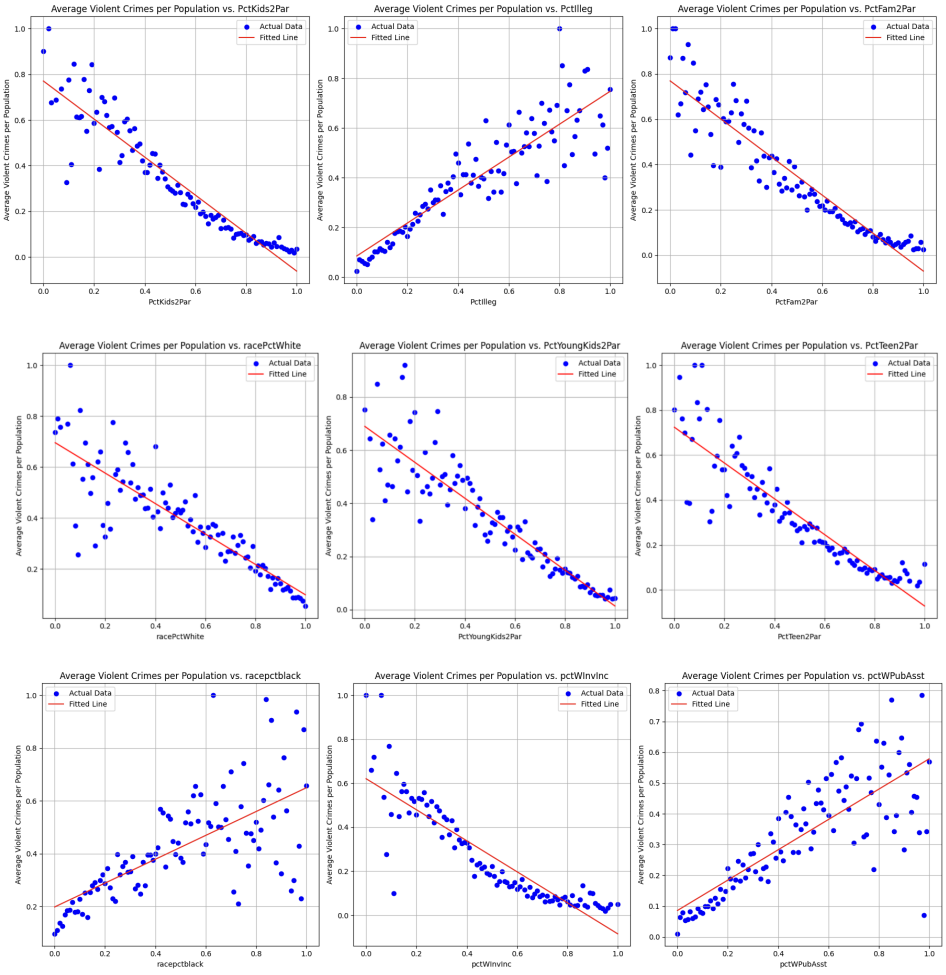
https://docs.google.com/document/d/1LbSPUtwteA5phU97Z-3GJhqpH-yved0_5GjnLDRGoys/edit?usp=sharing

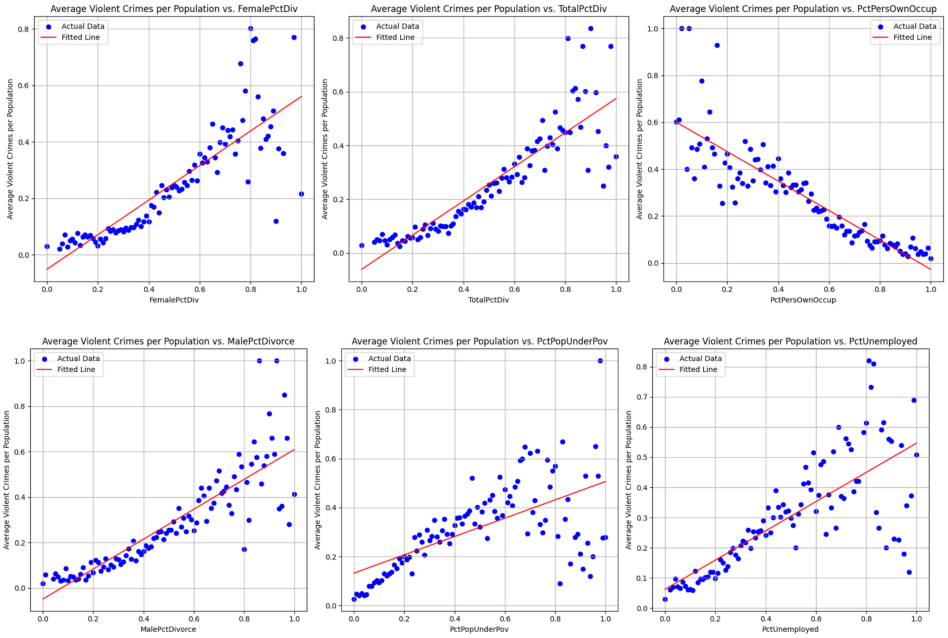
Appendix B




Violent Crimes per Pop vs. MalePctDivorce
Correlation: 0.53
Violent Crimes per Pop vs. PctPopUnderPov
Correlation: 0.52
Violent Crimes per Pop vs. PctUnemployed
Correlation: 0.50
Actual Data
Fitted Line
ViolentCrimesPerPop
MalePctDivorce
PctPopUnderPov
PctUnemployed

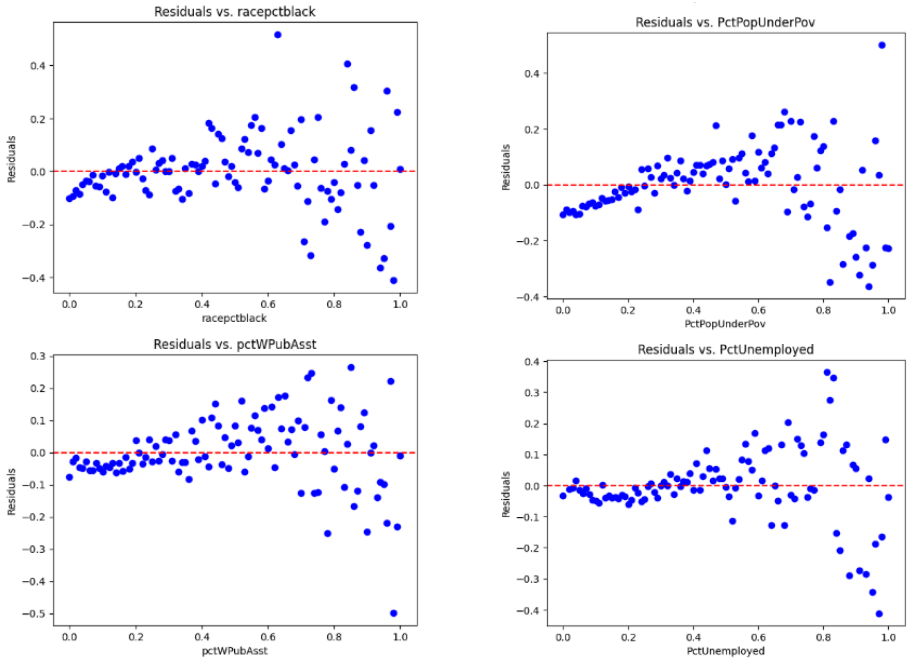
Appendix C


Average Violent Crimes per Population vs. PctKids2Par
Average Violent Crimes per Population vs. PctIlleg
Average Violent Crimes per Population vs. PctFam2Par
Average Violent Crimes per Population vs. racePctWhite
Average Violent Crimes per Population vs. PctYoungKids2Par
Average Violent Crimes per Population vs. PctTeen2Par
Average Violent Crimes per Population vs. racepctblack
Average Violent Crimes per Population vs. pctWInvInc
Average Violent Crimes per Population vs. pctWPubAsst
Actual Data
Fitted Line
Average Violent Crimes per Population
PctKids2Par
PctIlleg
PctFam2Par
racePctWhite
PctYoungKids2Par
PctTeen2Par
racepctblack
pctWInvInc
pctWPubAsst

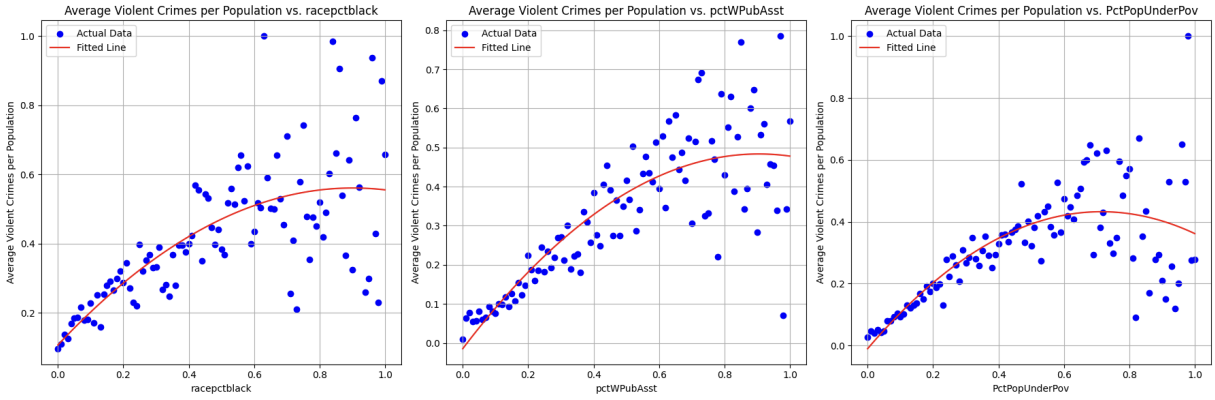


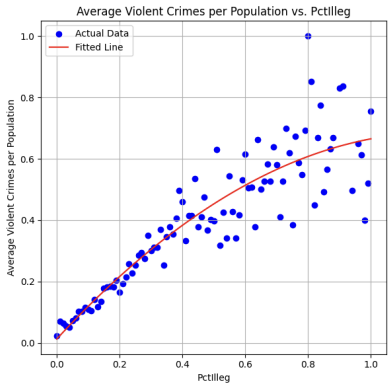
Feature	Correlation	P-Value
PctKids2Par	-0.882044194407537	0
PctIlleg	0.7837333400199356	0
PctFam2Par	-0.8824373377798281	0
racePctWhite	-0.7489992235068443	0
PctYoungKids2Par	-0.8187083452102637	0
PctTeen2Par	-0.8169728428623388	0
racepctblack	0.45470323320258665	0
pctWInvInc	-0.7808834760197372	0
pctWPubAsst	0.6244389870706466	0
FemalePctDiv	0.7118469194529691	0
TotalPctDiv	0.7795566203675028	0
PctPersOwnOccup	-0.7610812398378999	0
MalePctDivorce	0.7479657158286552	0
PctPopUnderPov	0.3835082996375565	0
PctUnemployed	0.5740063509020501	0

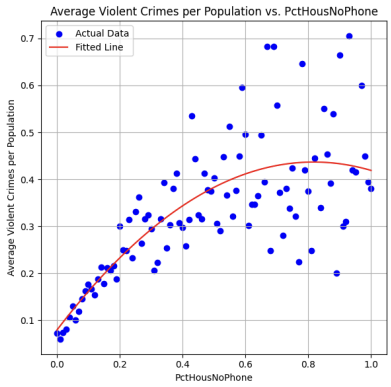
Appendix D - Residual Plot

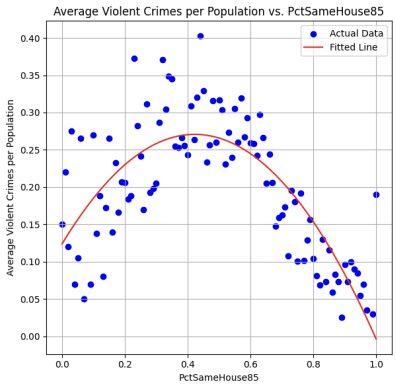


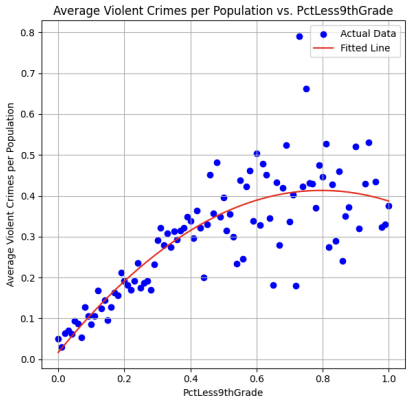
Appendix E

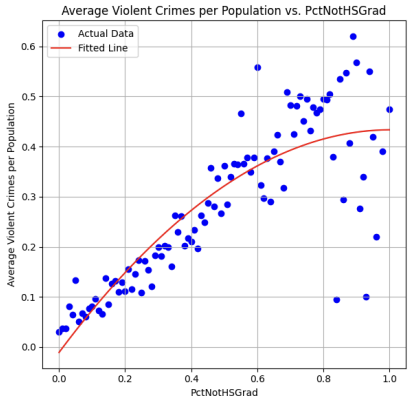


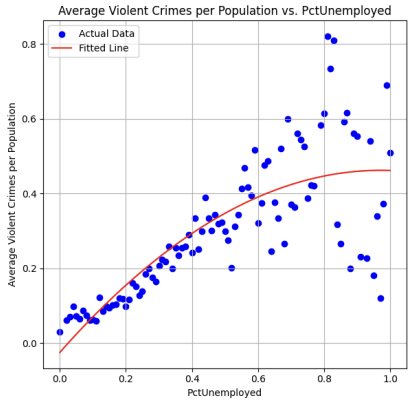


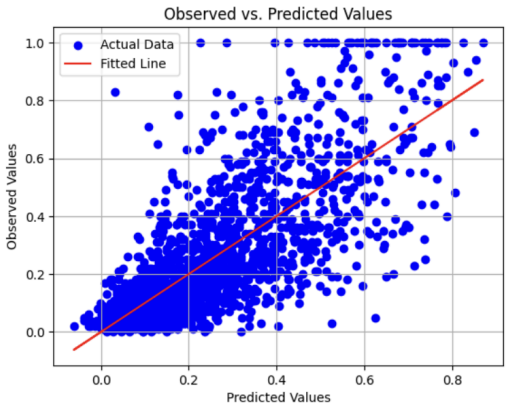












OLS Regression Results

Dep. Variable:	ViolentCrimesPerPop	R-squared:	0.595
Model:	OLS	Adj. R-squared:	0.591
Method:	Least Squares	F-statistic:	161.3
Date:	Tue, 14 May 2024	Prob (F-statistic):	0.00
Time:	16:47:20	Log-Likelihood:	977.56
No. Observations:	1994	AIC:	−1917.
Df Residuals:	1975	BIC:	−1811.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	−0.0063	0.026	−0.242	0.809	−0.057	0.044
PctIlleg	0.3332	0.077	4.330	0.000	0.182	0.484
PctHousNoPhone	−0.1969	0.076	−2.591	0.010	−0.346	−0.048
PctSameHouse85	0.1262	0.089	1.424	0.155	−0.048	0.300
PctUnemployed	0.0484	0.084	0.576	0.565	−0.116	0.213
PctNotHSGrad	0.0942	0.148	0.636	0.525	−0.196	0.385
PctLess9thGrade	−0.1967	0.128	−1.535	0.125	−0.448	0.055
PctPopUnderPov	0.5518	0.103	5.365	0.000	0.350	0.753
pctWPubAsst	−0.0019	0.091	−0.021	0.983	−0.181	0.177
racepctblack	0.4520	0.058	7.793	0.000	0.338	0.566
PctIlleg_sq	0.1047	0.071	1.469	0.142	−0.035	0.244
PctHousNoPhone_sq	0.1621	0.075	2.160	0.031	0.015	0.309
PctSameHouse85_sq	−0.1946	0.084	−2.321	0.020	−0.359	−0.030
PctUnemployed_sq	0.0233	0.084	0.279	0.780	−0.141	0.187
PctNotHSGrad_sq	0.1235	0.153	0.806	0.421	−0.177	0.424
PctLess9thGrade_sq	0.0346	0.125	0.278	0.781	−0.210	0.279
PctPopUnderPov_sq	−0.6205	0.095	−6.508	0.000	−0.807	−0.433
pctWPubAsst_sq	0.1581	0.084	1.878	0.061	−0.007	0.323
racepctblack_sq	−0.3256	0.061	−5.367	0.000	−0.445	−0.207

Omnibus:	325.839	Durbin-Watson:	1.939
Prob(Omnibus):	0.000	Jarque-Bera (JB):	891.372
Skew:	0.865	Prob(JB):	2.76e-194
Kurtosis:	5.781	Cond. No.	124.