Introduction

The goal of this project is to unravel the relationships between crime rates, violent crime rates, and key socio-economic factors — police funding and the educational profile of the town's residents with the help of Bayesian statistics. This information can help the town mayor decide on how to distribute the funds for the upcoming years.

Methods

Data Set

In this homework we used crime.csv data set, based on Life in America's Small Cities by G. S. Thomas. The data set reports crime rate and violent crime rate per 100,000 populants in small US towns, which will be our target variables. Furthermore, the data set also contains following features:

- *police_funding*: amount of money in millions of dollars that a particular town invested in police.
- *25_plus_high_school*: percentage of residents that are old 25 years or more and that completed a high school.
- *16_19_no_high_school*: percentage of residents aged between 16 and 19 that are not visiting a high school.
- *18_24_college*: percentage of residents aged between 18 and 24 that are enrolled in a college.
- *25_plus_4_years_college*: percentage of residents that are old 25 years or more and that visited a college for at least 4 years.

Data Analysis and Preparation

In Figure 1, the correlation matrix of the entire data set is presented. The highest correlation of 0.76 is observed between crime rate and violent crime rate. The next highest observed correlation is between *25_plus_high_school* and *25_plus_4_years_college*, amounting to 0.68. Additionally, it is noteworthy that among all features, the feature *25_plus_4_years_college* exhibits the lowest correlation with the target variables, crime rate, and violent crime rate. Due to its minimal correlation with the target variables and its substantial correlation with other education-related features, we have opted to exclude *25_plus_4_years_college* from the analysis. Consequently, the only remaining feature related to college education is *18_24_college*, which was renamed to just *college*. Furthermore, we have combined the features *25_plus_high_school* (feature 1) and *16_19_no_high_school* (feature 2) into a single feature named *high_school* using the following formula:

$$high\_school = 0.93 * feature1 + 0.07 * (100 - feature2) \qquad (1)$$

We calculated the weights based on publicly available demographic data for the U.S., indicating that individuals aged 25 and above are approximately 14 times more numerous than those aged between 16 and 19. The advantage of combining two features related to high school education into one is a reduced correlation between education-related features. A correlation of 0.21 was achieved between the *college* and *high_school* columns. Furthermore, min-max normalization was employed to scale each feature in a data set between a [0, 1] range, to ensure better interpretability of the model coefficients.

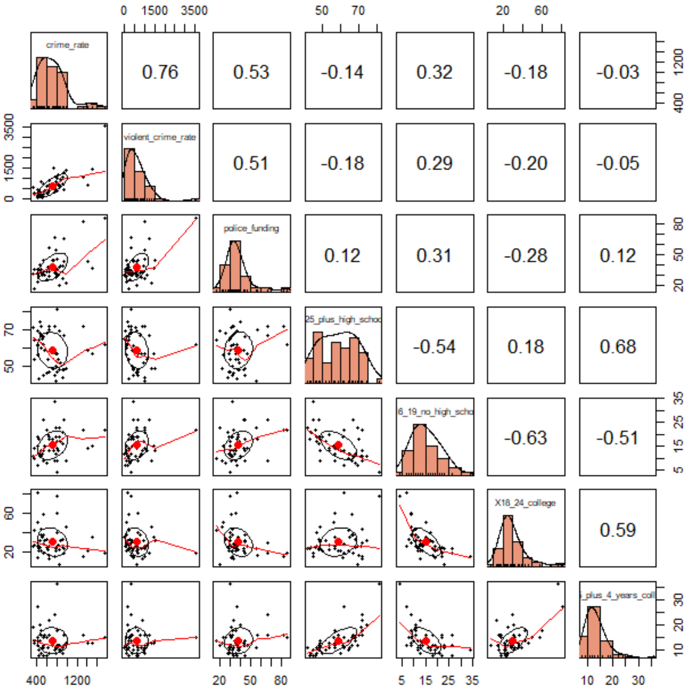

Figure 1. Correlation matrix of the original dataset

Model

Since the data for our target variables is non-negativite and right-skewned, we opted for the gamma regression models. Two gamma regression models were implemented, one for the *crime_rate* and one for the *violent_crime_rate*. The equations defining our models are:

$$\begin{aligned} y^{(i)} \mid \nu^{(i)}, \lambda^{(i)} &\sim \text{Gamma}(\nu^{(i)}, \lambda^{(i)}), \text{ where} \\ \nu^{(i)} &= \mu^{(i)} \lambda^{(i)} \\ \mu^{(i)} &= e^{\alpha^{(i)} + \beta_1^{(i)} * x_1 + \beta_2^{(i)} * x_1 + \beta_3^{(i)} * x_3} \\ \beta^{(i)} &\sim \text{Cauchy}(0, 2.5), \forall i \in \{\text{crime\_rate, violent\_crime\_rate}\} \end{aligned} \qquad (2)$$

The posterior distributions were explored using 4 MCMC chains, each with 1000 warm-up and 1000 sampling iterations.

MCMC Diagnostics

MCMC diagnostics, including Gelman-Rubin (R-hat), trace plots, and effective sample size, all indicated satisfactory results, affirming that MCMC chains have effectively explored parameter space and converged to the posterior distribution.

Results

The obtained results are presented in this section. The Figure 2. gives visual assessment of how well gamma regression models align with the actual distribution of the crime_rate and violent_crime_rate. We can observe that the model doesn't fit perfectly, but it suffices well for our analysis. Table 1 displays the average values and uncertainty of the coefficients for the crime rate and violent crime rate models, while their distributions are illustrated in Figure 3.

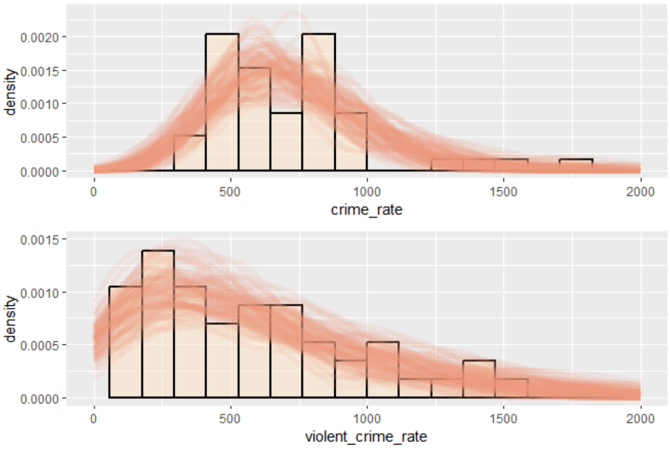

Figure 2. Original crime rate and violent crime rate data histograms overlaid with model predictions.

Table 1. Average Gamma regression coefficient values and their uncertainty.

	Crime Rate Model	Violent Crime Rate Model
intercept	6.38 ±0.13	6.26 ±0.27
police_funding	1.00 ±0.22	1.76 ±0.48
high_school	-0.36 ±0.18	-1.04 ±0.36
college	0.02 ±0.25	-0.17 ±0.52

From Table 1 we can see that the *police_funding* has the highest corresponding coefficient of all features for both crime rate and violent crime rate model, indicating that cities with higher police funding tend to have higher levels of criminal activity. This can be explained by the fact that cities with higher crime rates consequently need to allocate more funding to the police. The most negative coefficient for both models, -0.36 ± 0.18 for the crime rate and -1.04 ± 0.36 for the violent crime rate model, belongs to the column related to the percentage of people with high school education, suggesting that a higher percentage of high school educated residents reduces the frequency of criminal activities. These findings align with what we would expect based on the correlation matrix shown in Figure 1.

Furthermore, we observe that the *college* feature is accompanied by a negative average coefficient of -0.17 ± 0.52 in the violent crime rate model, while in the crime rate model, it is approximately zero (0.02 ± 0.25). Additionally, the coefficients associated with this feature are characterized by the highest uncertainty, indicating that the influence of this feature on the target variable in both models is surprisingly small.

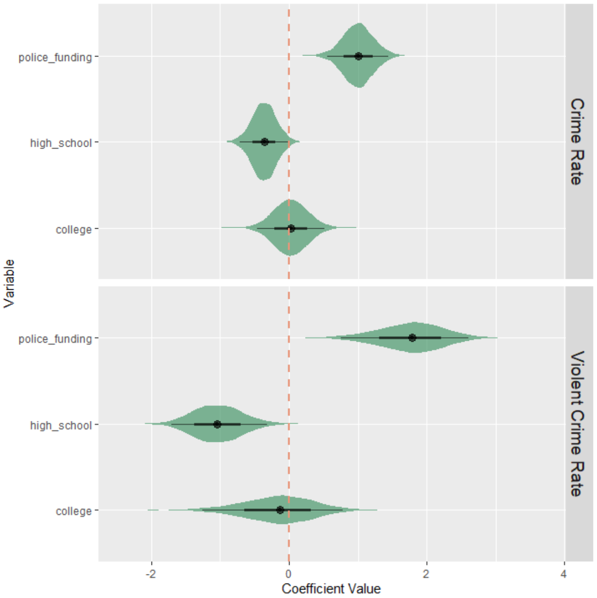

Figure 3. Distribution of the gamma regression coefficients for features of the crime rate and the violent crime rate model.

Discussion

The analysis results suggest that cities with more funding for the police tend to experience higher crime rates and violent crime rates. However, this doesn't mean that increased police funding directly causes more criminal activity. Instead, higher police funding might be a consequence of already elevated crime and violent crime rates in a given city. From the available data, it's not possible to conclude whether greater police funding can effectively reduce crime rates and violent crime rates.

Moreover, high school education has demonstrated a greater impact on reducing both crime rates and violent crime rates, compared to college education. The findings show that cities with a higher percentage of residents with high school education tend to have less criminal activities. This implies that investing in education, especially high school education, could contribute to a decrease in criminal activities.