Analyzing U.S. Crime Patterns Using Principal Component Analysis

Nathan Gates

2025-05-02

# 1 Executive Summary

The results of the study suggest that principal component analysis successfully reduced the dimensions of crime data across U.S. states. Two principal components explained most of the variation, allowing crime rates to be compared more easily across states and regions. Patterns of crime types and regional differences were identified through this analysis.

# 2 Introduction

The purpose of this study is to analyze state-level crime data from the United States for the year 2019. Seven different types of crimes were considered across all 50 states. Principal Component Analysis (PCA) was used to reduce the dimensionality of the data and uncover patterns across states and regions. Descriptive statistics and correlation analyses were also performed to summarize crime rates by state and by Census region.

## 2.1 Variables

Crime rates for seven different offenses were used in the study. The variables include rates per 100,000 people for each crime type. Each variable represents the number of reported offenses standardized by population.

Table 2.1: Variables used in the study

| Variable | Description |
| --- | --- |
| Burglary | Unlawful entry into a building to commit a crime. |
| Larceny | Theft of personal property without breaking and entering. |
| Car Theft | Theft or attempted theft of a motor vehicle. |
| Assault | Physical attack on another person. |
| Murder | Unlawful killing of another person. |
| Rape | Sexual assault against another person. |
| Robbery | Taking property from a person by force or threat. |

# 3 Methodology

Statistical analyses were conducted using RStudio. Descriptive statistics were used to summarize crime rates across states and regions. Principal Component Analysis (PCA) was then performed to reduce the number of variables and identify major patterns in the data.

## 3.1 Numerical Summaries

To begin, we computed summary statistics (mean and standard deviation) for each of the seven crime variables to understand the overall level and variability of crime across states. We also calculated these values by region to compare geographic differences. In addition, we examined the correlation matrix of the variables to explore how different crimes relate to one another. For instance, high correlations between burglary, larceny, and car theft would suggest these crimes tend to occur together.

## 3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used to reduce the number of correlated variables in a dataset while retaining as much information as possible. It does this by creating new uncorrelated variables, called principal components, which are linear combinations of the original variables. The general form of a principal component is:

where are the original standardized crime variables and are the loadings (coefficients) for each variable. Each principal component is constructed to maximize the amount of variance it explains while being uncorrelated with the other components. The loadings determine how much each original variable contributes to the principal component, and the scores represent the new coordinates of each observation in the reduced space.

# 4 4 Results

## 4.1 Descriptive Statistics Results

To understand the overall levels of crime across the country, we calculated the mean and standard deviation for each crime variable. Table 6.1 presents these values. Larceny had the highest national average, indicating it is the most common crime among the seven considered. In contrast, murder had the lowest average, which aligns with expectations. The standard deviations show large variability, especially for property crimes like burglary and larceny, suggesting that crime levels vary widely across states.

We next explored how these crimes relate to each other using a correlation matrix, which is displayed in Table 6.2. The matrix reveals strong positive correlations between property crimes such as burglary, larceny, and car theft, meaning that states with high rates of one of these crimes tend to also have high rates of the others. Violent crimes like assault and murder also show moderate positive correlations. Interestingly, rape stands out as the least correlated crime, indicating it behaves differently across the dataset.

Regional differences were also examined. Table 6.3 displays the mean and Table 6.4 display the standard deviation of each crime grouped by region. It is evident that the South and West regions tend to have higher average crime rates, particularly for burglary and larceny, while the Northeast consistently reports the lowest crime rates. The Midwest falls somewhere in the middle. These regional differences highlight how geography may play a role in crime trends.

## 4.2 Principal Component Analysis Results

To simplify the complexity of analyzing seven correlated variables, we applied Principal Component Analysis (PCA) to reduce the dimensionality. Figure 6.1 shows the scree plot of the eigenvalues. The sharp drop between the first and second components, followed by a leveling off, suggests that the first two principal components capture most of the variation in the data — specifically, 59.5% by PC1 and an additional 15.1% by PC2, totaling 74.6%. This suggests that plotting the data in two dimensions provides an effective summary of overall crime patterns.

We then examined how each crime contributes to the principal components. Table 6.5 presents the loadings, or weights, of each crime on the first two components. Most crimes load strongly and positively on PC1, indicating that this component represents overall crime severity. A state with a high score on PC1 is likely to have elevated levels of several crimes. PC2, on the other hand, separates crimes based on the nature of violent acts. Crimes like murder and robbery, which typically involve public confrontation and visible violence, have strong positive loadings, while rape, which often involves private, less visible acts of violence, has a strong negative loading. Therefore, PC2 can be interpreted as distinguishing between public confrontational violence and private personal violence. This separation highlights different patterns of violent crime among states.

The first two principal components are formed as follows:

The PCA biplot in Figure 6.2 shows how the 50 U.S. states are projected into the PCA space and colored by region. States in the Northeast cluster closely together, indicating relatively similar and lower crime rates. States in the South and West are more dispersed, suggesting greater variability and generally higher crime rates across different types.

The PCA variable plot in Figure 6.3 displays how each crime relates to one another in PCA space. This circular plot helps us see how crimes relate to each other. Crimes whose vectors are close to 0 degrees apart, such as burglary, larceny, and assault, are highly correlated and contribute similarly to PC1. In contrast, crimes whose vectors are close to 90 degrees, such as rape and robbery have no correlations with each other.

In summary, principal component analysis reduced the dimensionality of the crime data into two key components that capture most of the variation. The results highlight that property crimes tend to move together, while crimes like rape vary independently, and that regional differences play a role in crime rate patterns across the United States.

# 5 Conclusion

The principal component analysis was successful in reducing the complexity of crime rate data across U.S. states. It is our belief that this technique provides an effective way to summarize patterns among multiple crime types and detect regional differences. We strongly suggest that similar methods be applied to future crime datasets to better understand evolving crime trends across the country.

# 6 Appendix

Below are all of the graphs and tables used to conduct the statistical analysis.

Table 6.1: Mean and Standard Deviation of Crime Rates in the U.S.

| Crime | Mean | Standard Deviation |
| --- | --- | --- |
| Burglary | 347.30 | 142.39 |
| Larceny | 1,536.46 | 367.03 |
| Car Theft | 212.93 | 96.96 |
| Assault | 250.50 | 120.14 |
| Murder | 4.78 | 2.64 |
| Rape | 47.63 | 19.84 |
| Robbery | 62.58 | 32.33 |

Table 6.2: Correlation Matrix of Crime Variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Crime | Burglary | Larceny | Car Theft | Assault | Murder | Rape | Robbery |
| Burglary | 1.000 | 0.738 | 0.702 | 0.679 | 0.705 | 0.302 | 0.388 |
| Larceny | 0.738 | 1.000 | 0.778 | 0.614 | 0.588 | 0.342 | 0.462 |
| Car Theft | 0.702 | 0.778 | 1.000 | 0.571 | 0.388 | 0.442 | 0.483 |
| Assault | 0.679 | 0.614 | 0.571 | 1.000 | 0.660 | 0.546 | 0.494 |
| Murder | 0.705 | 0.588 | 0.388 | 0.660 | 1.000 | 0.168 | 0.588 |
| Rape | 0.302 | 0.342 | 0.442 | 0.546 | 0.168 | 1.000 | 0.072 |
| Robbery | 0.388 | 0.462 | 0.483 | 0.494 | 0.588 | 0.072 | 1.000 |

Table 6.3: Mean Crime Rates by Region

| Crime | Midwest | Northeast | South | West |
| --- | --- | --- | --- | --- |
| Burglary | 315.65 | 177.01 | 436.26 | 384.91 |
| Larceny | 1,436.00 | 1,078.81 | 1,725.52 | 1,713.32 |
| Car Theft | 205.22 | 92.28 | 219.71 | 295.23 |
| Assault | 239.68 | 145.64 | 291.17 | 283.04 |
| Murder | 4.13 | 2.69 | 6.98 | 4.11 |
| Rape | 51.34 | 34.59 | 41.40 | 60.91 |
| Robbery | 54.42 | 47.54 | 72.30 | 68.57 |

Table 6.4: Standard Deviations of Crime Rates by Region

| Crime | Midwest | Northeast | South | West |
| --- | --- | --- | --- | --- |
| Burglary | 60.36 | 28.35 | 147.61 | 129.78 |
| Larceny | 231.08 | 92.00 | 313.24 | 352.30 |
| Car Theft | 59.40 | 39.39 | 59.40 | 106.46 |
| Assault | 68.46 | 62.67 | 101.88 | 165.33 |
| Murder | 2.28 | 1.06 | 2.28 | 2.38 |
| Rape | 12.76 | 10.05 | 13.15 | 28.10 |
| Robbery | 24.73 | 28.12 | 29.29 | 41.31 |

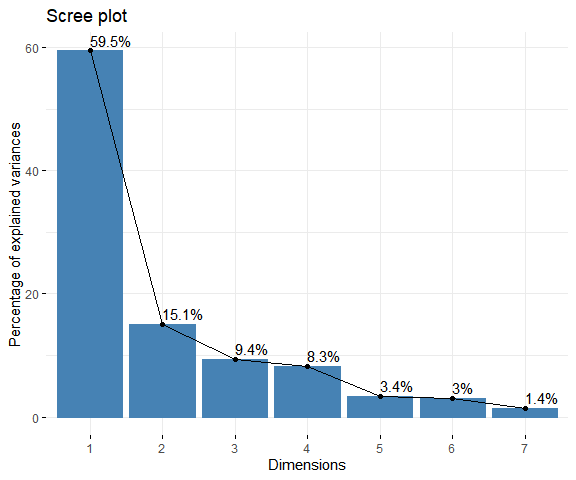


Figure 6.1: Scree plot showing the proportion of variance explained by each principal component.

Table 6.5: Principal Component Loadings for PC1 and PC2

| Crime | PC1 | PC2 |
| --- | --- | --- |
| Burglary | 0.425 | 0.041 |
| Larceny | 0.423 | -0.014 |
| Car Theft | 0.403 | -0.178 |
| Assault | 0.417 | -0.130 |
| Murder | 0.382 | 0.380 |
| Rape | 0.242 | -0.757 |
| Robbery | 0.317 | 0.482 |

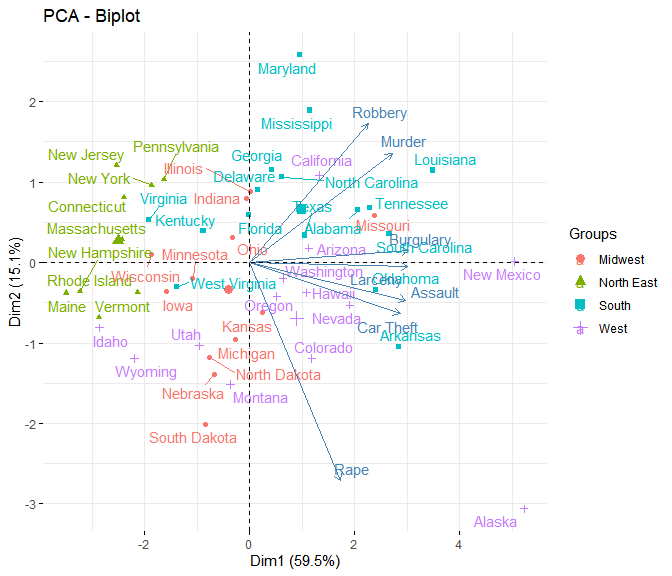


Figure 6.2: PCA biplot showing the states colored by region.

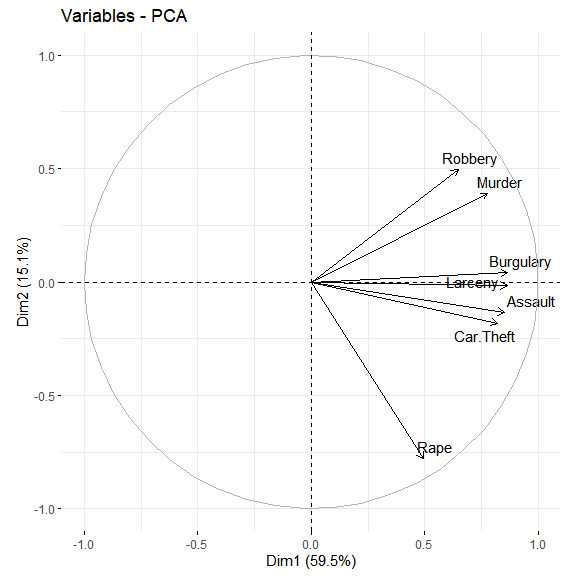


Figure 6.3: PCA variable plot showing relationships among crimes.