Middle East Technical University  
Department of Statistics

**— Final Report Project —**

**….s**Analysis of Crime Dynamics by Five States: Exploring Trends, Socioeconomic Factors, and Urbanization (2000-2023)

..  
**——————————**

by Burak Şahin Küçük, Fajrin Akbarli, Emek Bilge Aydın  
Furkan Karakurt, Anindya Haning Utami

2024 January | 2024 Fall

A red and white circle with a black background

Description automatically generated

# Abstract

This research study examines crime statistics from numerous states, concentrating on the relationship between crime rates and socioeconomic factors such as income, education, and urbanization. A dataset of 120 inputs is evaluated to identify trends and draw relevant conclusions that can potentially inform policy decisions. The research study sheds light on the links and correlations of crime dynamics through thorough data cleaning and the use of various visualization approaches. The study specifically includes findings on the distribution of crime rates between urban and suburban areas, the impact of income on property crime in connection to education levels, and a comparison of crime types by state. The findings highlight the intricate relationship between socioeconomic conditions and crime and provide a qualitative view of regional crime trends.

# Introduction

This report examines a comprehensive data set covering various aspects of crime-related statistics in different states. The focus is to analyze and understand the patterns and correlations between different socioeconomic factors and crime rates. Through this analysis, the report aims to uncover important trends and offer insights that can inform policymaking and societal understanding of crime dynamics. In this data, the relationship between the states in the dataset and crimes will be investigated with various research questions, and conclusions will be drawn from them.

## Data Description

The data collection under consideration has 120 items, each of which represents a unique set of observations on crime statistics in various states. The data initially featured issues such as missing values, errors, and formatting irregularities, which contributed to the data's disorganization. The table below shows the raw .data on the left side and the intended data on the right side to make it easier to comprehend.

**Table 1: Introducing data variables, making them understandable in corrected form**

|  |  |
| --- | --- |
| **Untidy Version** | **Clean Data Planned To Be Made** |
| **YEAR** (Continuous, but in this case, it is Discrete) | **Date** (int64 - integer) (The year of the observation.) |
| **STATE NAME** (object) | **State** (object) (The state in which the data was collected.) |
| **//Population** (Discrete) | **Population** (int64 - integer) (The population of the state.) |
| **Crime\_Type** (Object) | **Crime\_Type** (Object) (Type of crime) |
| **Arrests** (Discrete) | **Arrests** (int64 - integer) (Number of arrests the crime type) |
| **VICTIMS** (Discrete) | **Victims** (int64 - integer) (Number of victims affected by the crime.) |
| **Income** (Continuous) | **Income** (float64 - float) (Median household income in dollars.) |
| **Education\_Level** (Object) | **Education\_Level** (Object) (Highest education level achieved( |
| **Unemployment\_Rate** (Continuous) | **Unemployment\_Rate** (float64 - float) (Unemployment rate in the state.) |
| **!!Urbanization** (Categorical) | **Urbanization** (Object) (Level of urbanization) |
| **Number of property crimes reported** (Discrete) | **Property\_Crimes** (float64 - float) (Number of property crimes reported.) |
| **CRIME\_RATE** (Continuous) | **Crime\_Rate** (float64 - float) (Crime rate per 100,000 population.) |
|  | *(will be added)* **Arrests\_Rate** (float64 - float) (Arrests rate per 100.000 people in the state) |
|  | *(will be added)* **Victims\_Rate** (float64 - float) (Victims rate per 100.000 people in the state) |

Many variables in the data were used in the research questions. The removal of unused ones will take place in the data cleaning phase.

# Data Tidying and Cleaning Steps

Data must be successfully comprehended to be used to act, but incorrect data interpretation might lead to incorrect conclusions. Hence, some actions must be taken to understand data correctly. Data cleaning is a crucial step for drawing conclusions. Data cleaning was done entirely with Python libraries. Pandas was used to generally correct the dataset, while NumPy was used to fill in missing data. Some changes were made during the visualization phase. For example, while the values in the Income variable were left as integers, they were converted to salary format with periods and commas during the visualization phase. (94350 --> 94.350,00)

After introducing the data to pandas, the content of the data was initially evaluated with print statements to make sense of it. Functions such as.describe(),.info(), and.head() were used. The first two data lines were revealed to have nothing to do with the dataset. As a result, the first two lines were removed using the Pandas method (function). After these lines, the headers of the following line variables were made. Fixed misspellings in variable names; for example, //Population variable was corrected to Population. You can find all the new(changed) variable names in the table in the introduction.

It is important to modify the data types and incorporate the modifications into the master data. First, the .pd.to\_numeric method was used to alter the data type of variables having numeric data types. This means that the data specified in the table's right column in the introduction corresponds to the real data types (integer, float, object). Then, errors such as commas, periods, and spaces were repaired in the numerical data. For instance, the value "13148,13" has been changed to "13148.13" .unique function was used to verify for unique values in categorical variables. Misspelled words or phrases were discovered this way, and these errors were repaired using the .replace() and .title() methods. To clean the data, any gaps in category variables were removed using the .split() tool. Any duplicate data was deleted using the .drop\_duplicates() function while the data cleaning procedure was in progress. To detect missing data, the function is.null().sum() was utilized. This allowed us to see where the "NaN", gapped, or blank data were. We had four missing data points in total. We used the mode to fill in the missing categorical data and the mean to fill in the missing numerical data. The mean was utilized when the difference between the mean and the median was minor. The median was preferred if the value differences in sample size were large. Considering that the number of observations was already small, and the data was not much affected by these fillings. We did not delete or remove any cases because of these considerations.

In addition, an outlier detection function was developed. By running this procedure, several outliers were discovered in the Unemployment rate column, but because this column is critical to us, we did not remove or delete any outlier observations. As a result of this "Unemployment\_Rate", we have two outliers. In the other numerical variables, there are no outliers. Some mistakes in education level were discovered. The "Education\_Level" values for the states were nearly identical every year, but some "Date" had minor inaccuracies. As a result, states and their education levels were matched, and minor adjustments or inaccuracies were fixed. In the same way, the "Urbanization" values were mapped to the state values they should be. Mistakes fixed. The information was well-organized. The data was then transferred from Pandas to Excel. When comparing the findings, no errors, misinterpretations, missing values, or other issues were discovered.

# Exploratory Data Analysis

## Evaluating Variables

It is necessary to create and infer visualizations to make meaning of this data set effectively. For humans, pure numerical data is meaningless. Therefore, five research questions were identified to make sense of the data and to draw conclusions. Matplotlib and Seaborn were used to visualize the graphs for these questions. Furthermore, guidance was given to put NumPy to use in determining the correlation coefficient for the specific question.

## Descriptive Statistics

## Research Questions

### How Has the Overall Violent Crime Rate Changed From 2000 to 2023 in total 5 states?

A graph with a line going up

Description automatically generated   
**Figure 1: Line Chart of Overall Violent Crime Rate From 2000 to 2023**

Given Figure 1\*, the graph provides a clear visual depiction of the changes in violent crime rates over the indicated time period. The line graph, marked with red dots connected by lines, shows the trend in violent crime rates per 100,000 people in the five states. The graph shows an initial decline in violent crime rates since 2000 and then fluctuates over the years. Although there are noticeable ups and downs, a clear upward trend emerges from 2014 onwards. The shaded area around the line indicates a range of uncertainty, indicating variability and probability within each year. The lowest crime rate was recorded in 2005. The highest crime rate occurred in 2023. Although the things that we cannot define now, which were applied until 2005, caused the crime rate to dip, the effect of these things that we cannot define was removed, and crime rates gradually increased.

The years from 2011 to 2023 show an increasing trend, with the crime rate rising more sharply towards the last years. This may indicate various social, economic, or other factors influencing the increase in violent crime. This visualization provides an important tool for understanding the evolution of violent crime rates and highlights the need for in-depth analysis to identify the underlying causes of these trends. This information can be vital for law enforcement agencies to develop strategies and take effective measures to address the increase in violent crime rates.

### How does the distribution of crime rates differ between urban and suburban areas?

A screenshot of a graph showing a leaf

Description automatically generated  
**Figure 2: Line Chart of Overall Crime Rate From 2000 to 2023**

Given Figure 2\*, the violin plot effectively encapsulates the nuanced patterns and variations in crime rates across different years between urban and suburban areas. The width of each violin provides a visual representation of the interquartile range: 84.5 for the urban plot and 65.55 for the suburban plot, with the median for each being 156.7 and 136.9 for the crime rate denoted by the central point. This allows a quick comparison of the central tendency and variability in crime rates within each urban or suburban category. The violins' shapes illustrate the data distribution, with more comprehensive sections indicating higher variability, while narrower portions suggest more consistency in crime rates. Outliers within any data point below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR. Any data point below 15.45 or above 353.45 in the urban violin plot and any data point below 17.575 or above 279.775, represented as individual data points beyond the whiskers, draw attention to exceptional high or low crime rates, offering valuable insights into potential anomalies or noteworthy trends over the years.

Examining specific statistical measures within the violin plot enhances the interpretative depth. Including quartiles such as the first quartile (urban: 142.2 and suburban: 115.9 and third quartile (urban is 226.7 and suburban is 181.45) helps delineate the spread of crime rates. At the same time, the maximum and minimum values provide a clear understanding of the data range. These details contribute to a comprehensive analysis, enabling a more informed understanding of how crime rates have evolved and differed between urban and suburban settings while pinpointing periods with notable deviations from the norm.

### How Does the Rate of Victimization Differ Among Various States?

A graph showing a number of different colored squares

Description automatically generated  
**Figure 3: Five Box Plots for Comparison Victims Rate Distribution by State**

The box plot in Figure 3\* provides a visual comparison of the rates at which people are affected by crime as victims. In five states: California, Texas, Florida, New York, and Pennsylvania. The rate is expressed in terms of 100,000 people. A box representing a state depicts each state's victimization rate distribution. The median rate, or the midpoint of all observed rates, is represented by the line in the center of each box. According to this information, the Median of California is almost 440, Texas is almost 600, Florida is almost 720, New York is almost 670, and Pennsylvania is almost 850 per 100.000 people. The highest median value is in Pennsylvania, and the lowest median value is in California for Victims Rate. Half of the observed rates fall inside the box, represented by the ends corresponding to the upper and lower quartiles. The "whiskers" that protrude from the boxes display the data range, omitting outliers. The values of Q1 and Q3, respectively, for each state are approximately as follows. California is 280 and 680, Texas is 380 and 895, Florida is 410 and 1025, New York is 370 and 1045, Pennsylvania is 440 and 1680. Compared to other states, Pennsylvania's victimization rate appears to vary more, as evidenced by the visualization's wider box. Texas and California, on the other hand, have smaller boxes, suggesting a more narrow range of rates.

Additionally, Pennsylvania's box's median line is higher than other states, indicating that the state's victimization rate is likely higher on average. This may suggest that further investigation is required to fully understand the reasons behind the crime and possible preventative measures in Pennsylvania. We can learn more about the average victimization rate and the consistency or variability of the rates in each state by comparing them.

### How Does Income Level Correlate With the Incidence of Property Crimes Across Different Education Levels?

A graph showing the difference between income and property crime

Description automatically generated  
**Figure 4: Scatter Plot which Indicates that Correlation Between Income vs Property Crimes with Education Levels**

The graph in Figure 4\* provides a visual examination of the relationship between income levels and the occurrence of property crimes, with an additional focus on different education levels. The scatter plot shows two groups of points, each colored differently to represent the two education levels: green for "High School" and blue for "College". Also, the correlation coefficient "r" = 0.93, which indicates a strong positive correlation, supported by a line of best fit that passes through the data points. This value, which is almost near 1, points to a clear correlation between property crimes and income. In this context, it indicates a tendency for the number of property crimes to rise with income.

According to the plot, there is an observable association between higher income brackets and an increased prevalence of property crimes among individuals with higher education levels, specifically those who have attended college. Conversely, individuals who have completed high school demonstrate a narrower income distribution and a lower frequency of property crimes.

The data suggests a potential correlation between higher income levels, elevated educational attainment, and an increased occurrence of property crimes. Contributing factors may include the higher value of local real estate or elevated reporting of crimes in affluent areas. The analytical findings emphasize the importance of considering economic and educational factors when devising crime prevention strategies, providing valuable insights for law enforcement and community planners. Notably, income was initially not converted to dollars during the data-cleaning phase, given its versatile allocation. However, for the specific analysis, chatbots were utilized to convert income to dollar format, ensuring a standardized and comparable representation that enhances the clarity and interpretability of the results.

### What is The Relationship Between Crime Type and Overall Crime Rates in Each State?

**Figure 5: Heat Map Highlights Crime Types vs Overall Crime Rates by State**

A screenshot of a graph

Description automatically generatedThis given Visualization 5\* the heatmap visually represents the average crime rates for two types of crimes—property and violent—across five different states: California, Florida, New York, Pennsylvania, and Texas. The color intensity corresponds to the average crime rate, with the scale shown on the right side of the heatmap. Darker shades represent higher crime rates, while lighter shades indicate lower crime rates. California has the highest rate of violent crimes among the states listed, as indicated by the darkest shade in the 'Violent' column. Florida has a notably lower rate of violent crimes compared to its property crimes, which is the opposite pattern of what is observed in California. New York has nearly similar rates for both crime types, as the colors for both categories are quite similar. Texas shows the lowest average rate for violent crimes, yet its rate for property crimes is slightly higher than for violent crimes. This heatmap provides a clear and immediate understanding of how crime rates differ by type across these states, which can be crucial for law enforcement agencies and policymakers to allocate resources and implement targeted crime prevention strategies effectively.

# Conclusion

Consequently, the information from this report should be used to tailor crime prevention and intervention strategies that consider the socioeconomic background of each individual state. While these correlations exist, it is imperative to recognize that they do not imply causality, and a deeper reflection on causal factors is recommended for a more comprehensive understanding. The data also underscores the importance of continuous monitoring and analysis to adapt to the changing nature of crime and the factors contributing to crime. The conclusions drawn from this data inform us of the fact that California as a state has the highest number of crimes to many other factors with thanks to research questions.

# Sources and Documents

Many steps are planned to make this data meaningful. More than 5 applications and Python libraries were used for these stages. Tasks were distributed among the group members. Cloud files were created to manage this operation. We also got help from the internet for color palettes. Here, we thank McKinsey and Dribbble freelancers. Below are all our source files, pure data, edited data, and the applications we used. In the reference section, the sites of the owners of the color palettes we used are placed. It has full regulatory authority over these links. Please do not share links with anyone unless you have to. Links except GitHub will be destroyed 2 months after the report (60 days - March 18, 2024)

Burak Şahin Küçük: 2614758  
Fajrin Akbarli: 2656072  
Anindya Haning Utami: 2602910  
Emek Bilge Aydın: 2614444  
Furkan Karakurt: 2614691

**References**

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

McKinney, W., & others. (2010). Data structures for statistical computing in python*. In Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

*Meet the new McKinsey look*. (2019b, February 24). McKinsey & Company. <https://www.mckinsey.com/about-us/new-at-mckinsey-blog/meet-the-new-mckinsey-look>

*Faiz 🍉*. (n.d.). Dribbble. <https://dribbble.com/faizz_anwar>

Python Software Foundation*.* (n.d.). *Python Language Reference, version 3.8. Retrieved from* [*https://www.python.org*](https://www.python.org)

Waskom, M. L., (2021).seaborn: statistical data visualization*. Journal of Open Source Software, 6(60), 3021,* [*https://doi.org/10.21105/joss.03021*](https://doi.org/10.21105/joss.03021)*.*

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. (*[*Publisher link*](https://www.nature.com/articles/s41586-020-2649-2)*).*

Google Colaboratory. (n.d.). https://colab.research.google.com/