Navid Hoque

CMPSC 463

12/3/2024

Project 2

1.Project Goals:

The primary objective of this project is to analyze and visualize crime data across US states using statistical and machine learning methods. The analysis focuses on identifying patterns, relationships, and anomalies in crime rates (Murder, Assault, Rape) and their correlation with the urban population.

The specific goals include:

1. **Visualizing Crime Metrics**:
    - Generate clear and comprehensive visualizations for each crime type (Murder, Assault, Rape, UrbanPop) to understand state-wise distributions and trends.
2. **Analyzing Relationships**:
    - Investigate the correlation between the percentage of urban population and individual crime rates.
    - Use scatterplots and regression analysis to quantify and visualize these relationships.
3. **Identifying Anomalies**:
    - Detect states with unusual crime rates or urban population percentages through z-score calculations and highlight them in visualizations.
4. **Regression Analysis**:
    - Fit regression models to explore linear relationships between urban population percentages and crime rates.
    - Provide insights into the direction, strength, and significance of these relationships.
5. **Actionable Insights**:
    - Deliver meaningful insights into crime patterns and outliers that could inform policymakers and stakeholders.

2. Significance and novelty of the project

Background Information

Understanding crime patterns is a critical component of effective policymaking, law enforcement, and public safety planning. Urbanization has long been suspected to influence crime rates, but quantifying and visualizing these relationships remains a challenge due to the complexity of socio-economic and demographic factors. This project leverages state-wise crime data (Murder, Assault, Rape) and urban population statistics to explore these dynamics.

Why This Project is Meaningful

1. **Insights for Decision-Makers**:
   - The project provides policymakers and law enforcement agencies with actionable insights by highlighting correlations between urbanization and specific crimes. This can help in resource allocation, urban planning, and crime prevention strategies.
2. **Data-Driven Analysis**:
   - By combining statistical techniques (correlation, z-score) with regression modeling, the project offers a data-driven approach to understanding the factors influencing crime rates.
3. **Visual Representation of Trends**:
   - The use of individual visualizations, scatterplots, and regression lines makes complex data accessible to a wider audience, bridging the gap between raw data and actionable insights.
4. **Anomaly Detection**:
   - Identifying states with unusually high or low crime rates compared to their urban population provides a deeper understanding of regional factors that deviate from general trends.

Novelty of the Project

1. **Integration of Urbanization as a Factor**:
   - Unlike many crime analysis projects that focus solely on crime rates, this project introduces urban population as a key variable, providing a fresh perspective on the urbanization-crime dynamic.
2. **End-to-End Approach**:
   - From data visualization to anomaly detection and regression modeling, the project offers a comprehensive framework for analyzing crime data.
3. **Anomaly Highlighting in Visualizations**:

   - States with anomalous crime rates are explicitly labeled on scatterplots, offering an intuitive way to identify outliers that deviate from expected trends.

4. **Actionable Metrics**:
   - Regression coefficients and anomalies are calculated and visualized, making the results directly usable for stakeholders looking to address crime in urbanized areas.

This project stands out for its integration of statistical rigor, novel use of urbanization data, and emphasis on making results accessible and actionable. Let me know if you'd like additional points or refinements!

3. Installation and usage instructions:

1. Clone or download the Project Repository
2. Set up python environment
3. Install required Packages
   a. Matplotlib
   b. Pandas
   c. Seaborn
   d. Scikit-learn
   e. Numpy
   f. Scipy

Usage Instructions:

1. Place the cleaned dataset in the project directory
2. Run the Crime_DataVisualization.py file
3. The outputs will be many different bar charts and scatter plots first showing total crime by state then to scatter plots of individual crime by urban population

4.Code Structure:

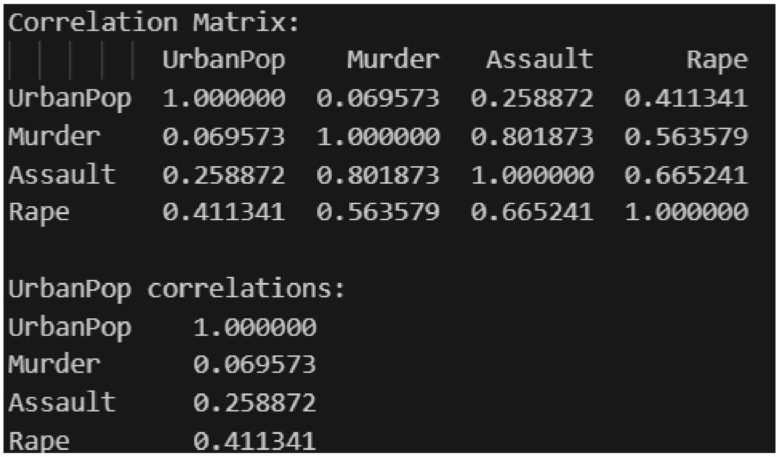
- **Block Diagram:** The structure of the code can be represented by the following flowchart:

```
+---------------------------+
| Start: Load CSV Dataset   |
+---------------------------+
          |
          v
+---------------------------+
| Preprocess Data:          |
| - Define crime_columns    |
+---------------------------+
          |
          v
+---------------------------+
| Step 1: Visualizations    |
| - Bar Charts           |
| - Scatterplots         |
+---------------------------+
          |
          v
+---------------------------+
| Step 2: Correlation      |
| - Compute Correlation    |
| - Print Correlation Matrix|
+---------------------------+
          |
          v
+---------------------------+
| Step 3: Anomaly Detection |
| - Z-Score Calculation    |
| - Identify Anomalies     |
| - Highlight Anomalies    |
+---------------------------+
          |
          v
+---------------------------+
| Step 4: Regression Line  |
| - Fit Linear Model       |
| - Plot Regression Line   |
| - Output Coefficients    |
+---------------------------+
          |
          v
+---------------------------+
| End: Generate Insights   |
| - Visualizations         |
| - Regression Analysis    |
+---------------------------+
```

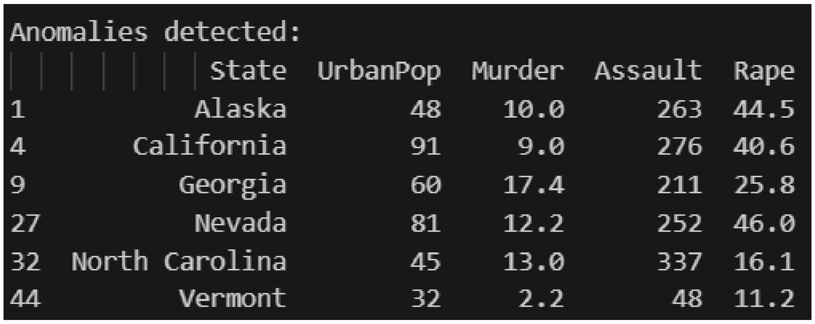
Code explanations:

1. Loading the dataset
    a. UScrime_cleaned.csv is loaded using pandas
    b. Crime columns were labeled as Murder, Assault, Rape, and UrbanPop (I understand Urban population is not a crime but to show representation of it per state I had to put it there)
2. Visualizations
    a. Bar charts
        i. Created individual bar charts for each “crime” type across all states

        ii. Provides a state wise breakdown of each crime metric
    b. Scatterplots:
        i. Visualizes the relationship between Urban Population (UrbanPop) and each crime type

3. Correlation Analysis
    a. Calculates the correlation matrix to identify relationships between UrbanPop and crimes
    b. Prints correlation values highlighting the strength and direction of relationships
        i. 

```
Correlation Matrix:
          UrbanPop    Murder   Assault      Rape
UrbanPop  1.000000  0.069573  0.258872  0.411341
Murder    0.069573  1.000000  0.801873  0.563579
Assault   0.258872  0.801873  1.000000  0.665241
Rape      0.411341  0.563579  0.665241  1.000000

UrbanPop correlations:
UrbanPop    1.000000
Murder      0.069573
Assault     0.258872
Rape        0.411341
```

4. Anomaly Detection
    a. Z-Score Calculation:
        i. Computes z-scores for each metric to standardize the data
    b. Identify Anomalies
        i. Detects states with unusually high or low z-score values
    c. Highlight Anomalies
        i. Anomalies are labeled directly on scatterplots for clear identification
        ii. 

```
Anomalies detected:
             State  UrbanPop  Murder  Assault  Rape
1           Alaska        48    10.0      263  44.5
4       California        91     9.0      276  40.6
9          Georgia        60    17.4      211  25.8
27          Nevada        81    12.2      252  46.0
32  North Carolina        45    13.0      337  16.1
44         Vermont        32     2.2       48  11.2
```

        iii. Also listed as shown above

5. Regression Line
    a. Fits a linear regression model to quantify the relationship between UrbanPop and each crime
    b. Plots the regression line over scatterplots
    c. Outputs regression coefficients and intercepts for interpretation

5.List of Functionalities and Verification results:

Functionalities:

1.Data Loading and Preprocessing:

- **Functionality**: Loads the cleaned dataset UScrime_cleaned.csv using pandas.
- **Verification**: The dataset is correctly loaded into a DataFrame, and columns for analysis (Murder, Assault, UrbanPop, Rape) are identified.
- **Test Result**: Verified by printing crime_data.head() to ensure correct column names and data structure.

Bar Chart Visualizations for Each Crime:

- **Functionality**: Generates individual bar charts showing state-wise crime rates for Murder, Assault, UrbanPop, and Rape.
- **Verification**: Visualizations correctly display crime metrics for all states with appropriately labeled axes and titles.
- **Test Result**: Verified by visual inspection of the generated charts.
-

Scatterplots of Urban Population vs Crimes:

- **Functionality**: Plots scatterplots for UrbanPop against Murder, Assault, and Rape to observe relationships.
- **Verification**: Scatterplots accurately plot UrbanPop on the x-axis and the selected crime metric on the y-axis, showing expected trends and variations.
- **Test Result**: Verified by observing clear relationships or lack thereof in the plots.

Correlation Analysis:

- **Functionality**: Computes and prints the correlation matrix for UrbanPop and the crime metrics.
- **Verification**: Correlation values align with observed trends in scatterplots (e.g., positive correlation indicates that higher UrbanPop correlates with higher crime rates).
- **Test Result**: Verified by matching scatterplot observations with correlation values.

Anomaly Detection:

- **Functionality**:
    - Calculates z-scores for UrbanPop and crime metrics to standardize data.
    - Identifies anomalies with z-scores greater than ±2 and highlights them on scatterplots.

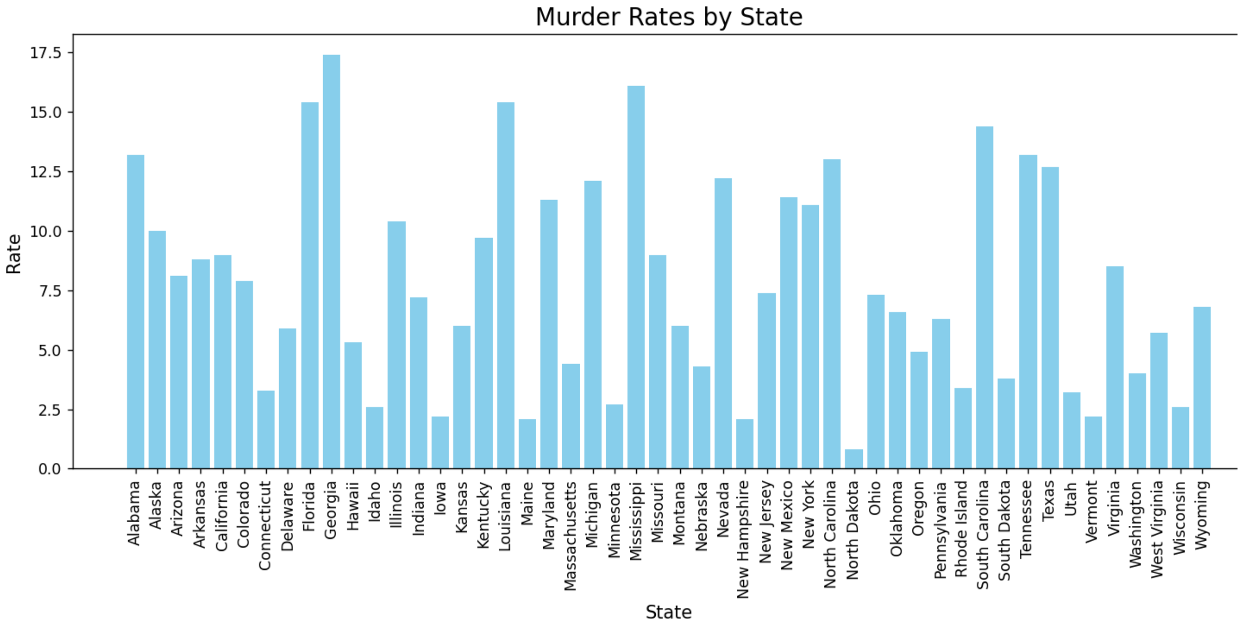
- **Verification**:
    - Verified by printing detected anomalies.
    - Anomalous states are correctly labeled on scatterplot

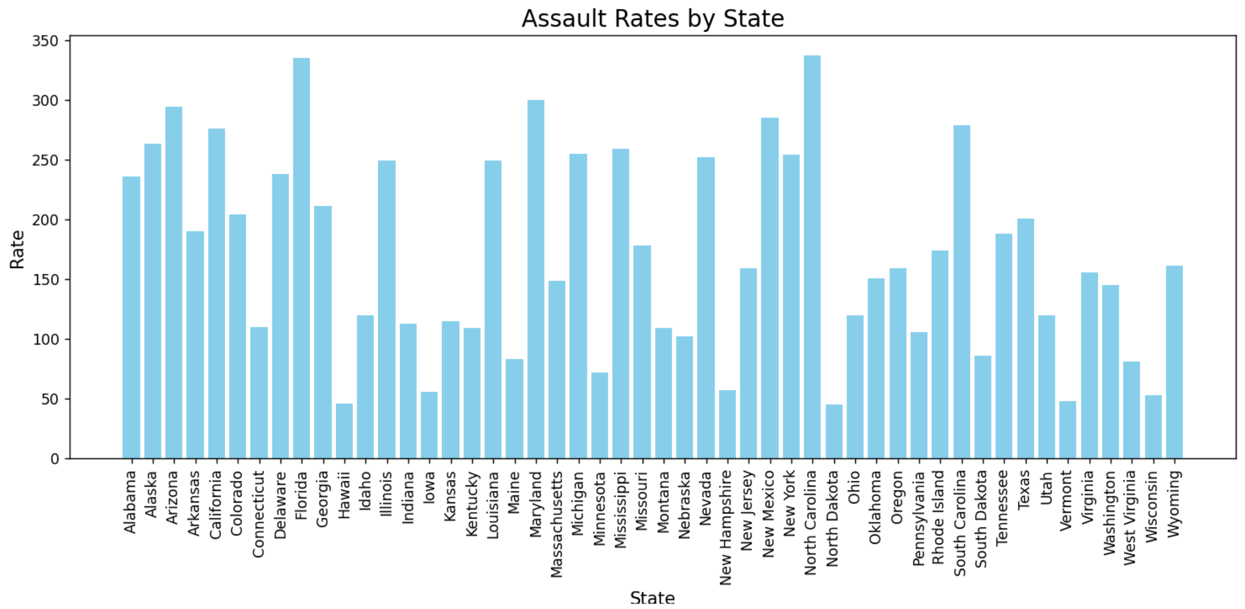
Regression Analysis:

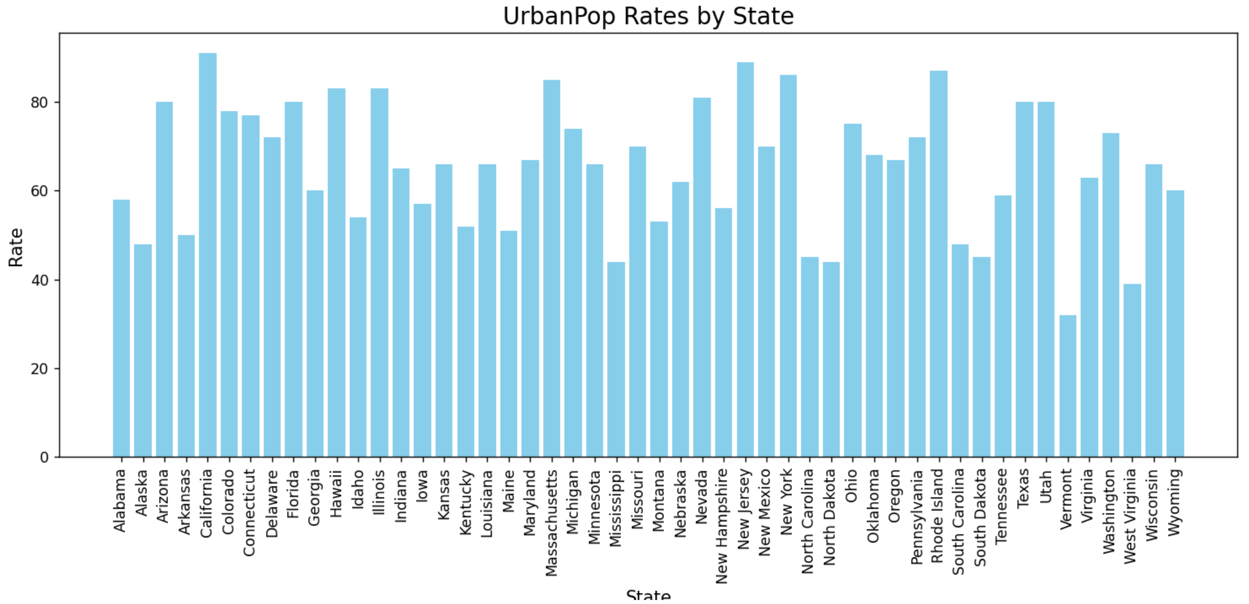
- **Functionality**:
    - Fits a linear regression model to quantify the relationship between UrbanPop and each crime metric.
    - Plots regression lines on scatterplots to visualize trends.
    - Outputs regression coefficients and intercepts for interpretation.
- **Verification**:
    - Regression lines align well with scatterplot data points.
    - Coefficients are meaningful and match observed trends.

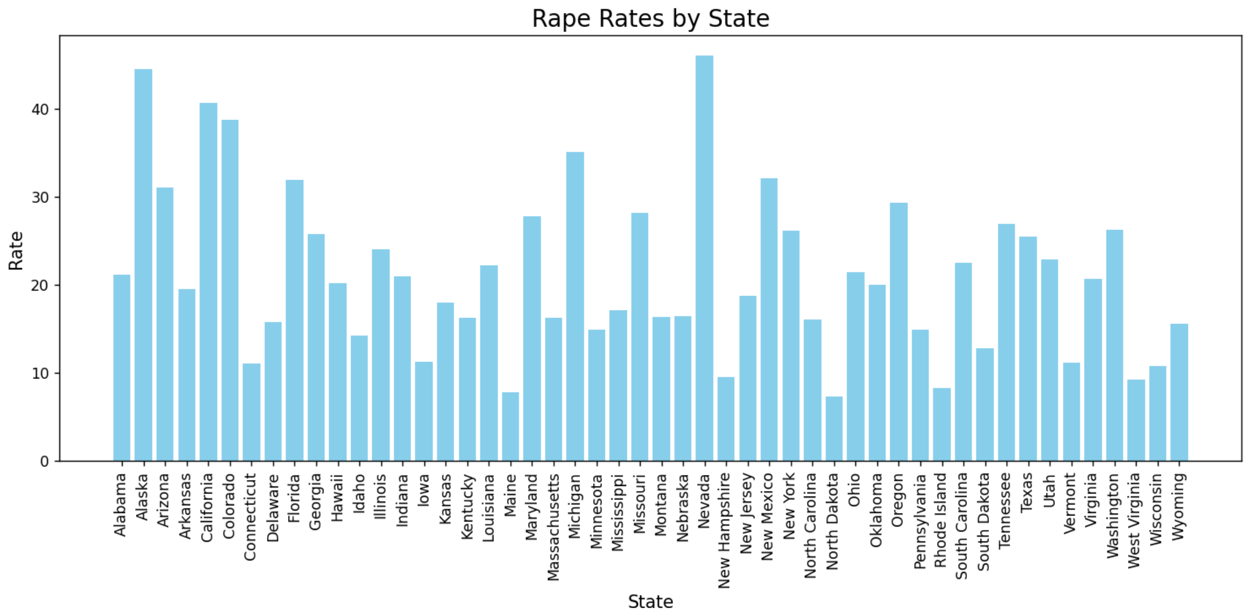
6.Showcasing the Achievement of Project Goals:

Bar graphs of Crime and Urban Population by State

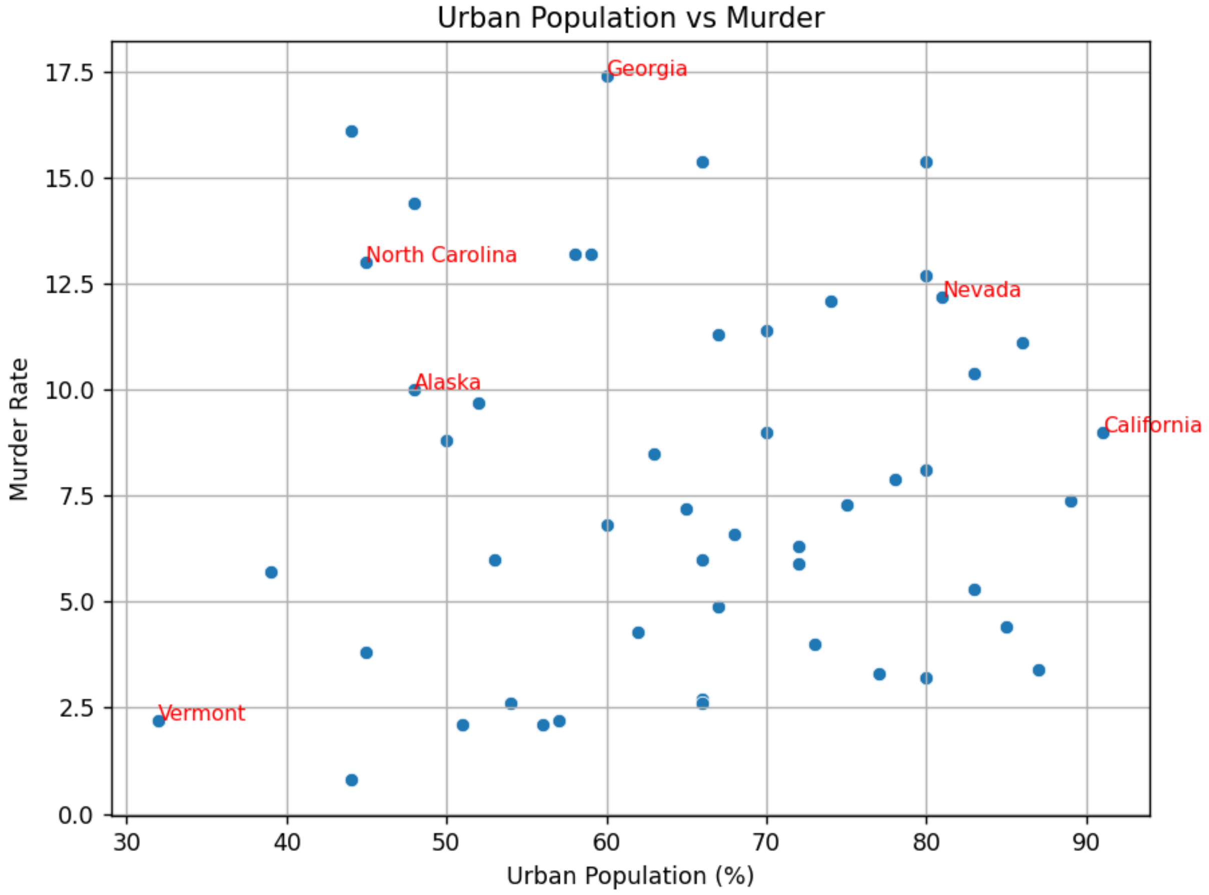


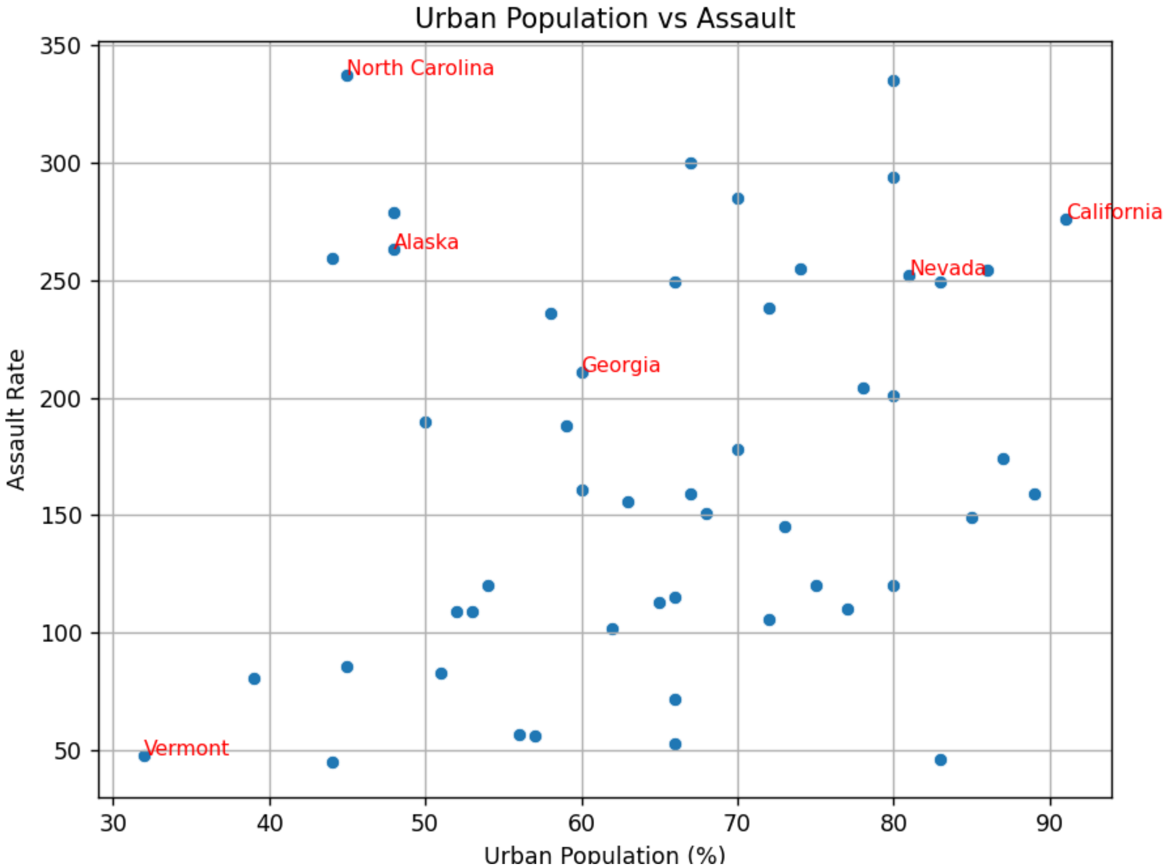

Assault Rates by State
Rate
0
50
100
150
200
250
300
350
State
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

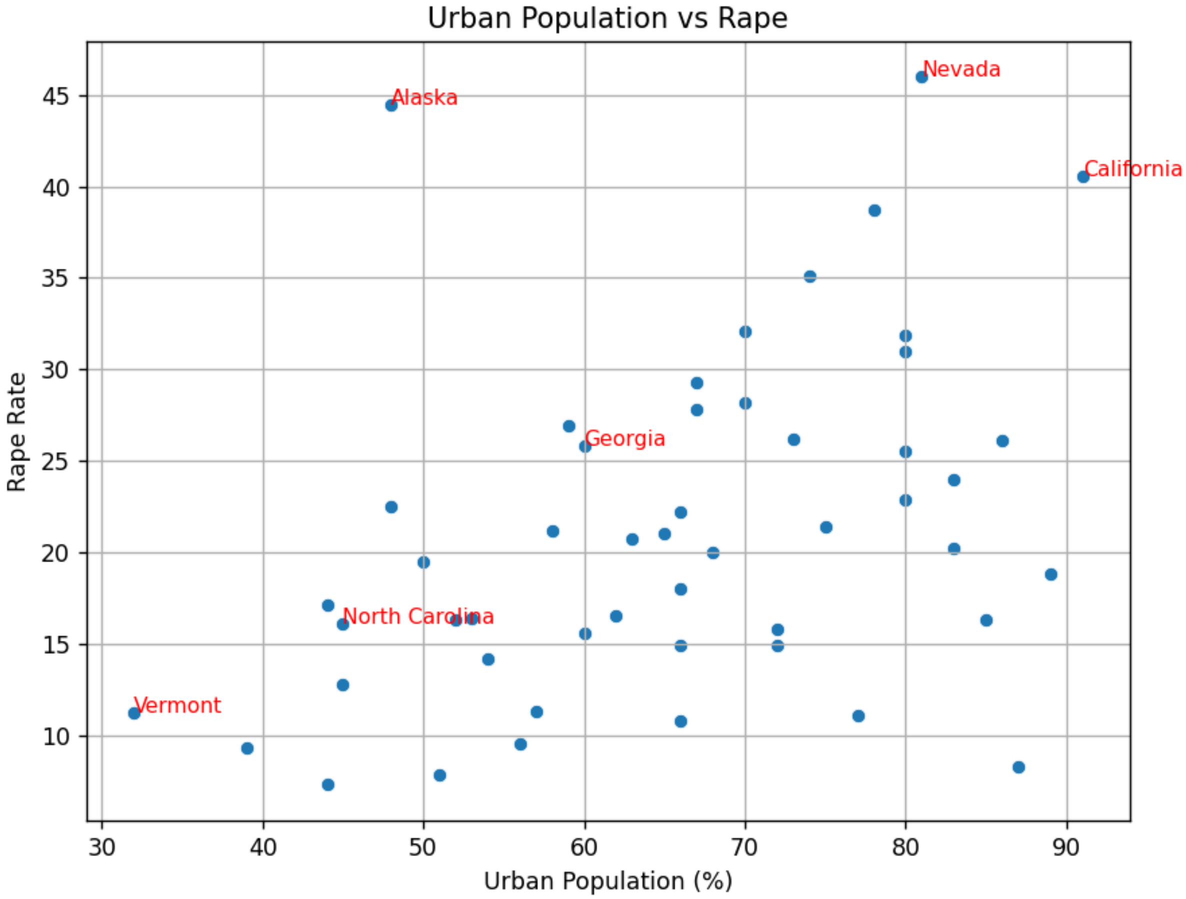

UrbanPop Rates by State
Rate
0
20
40
60
80
State
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming



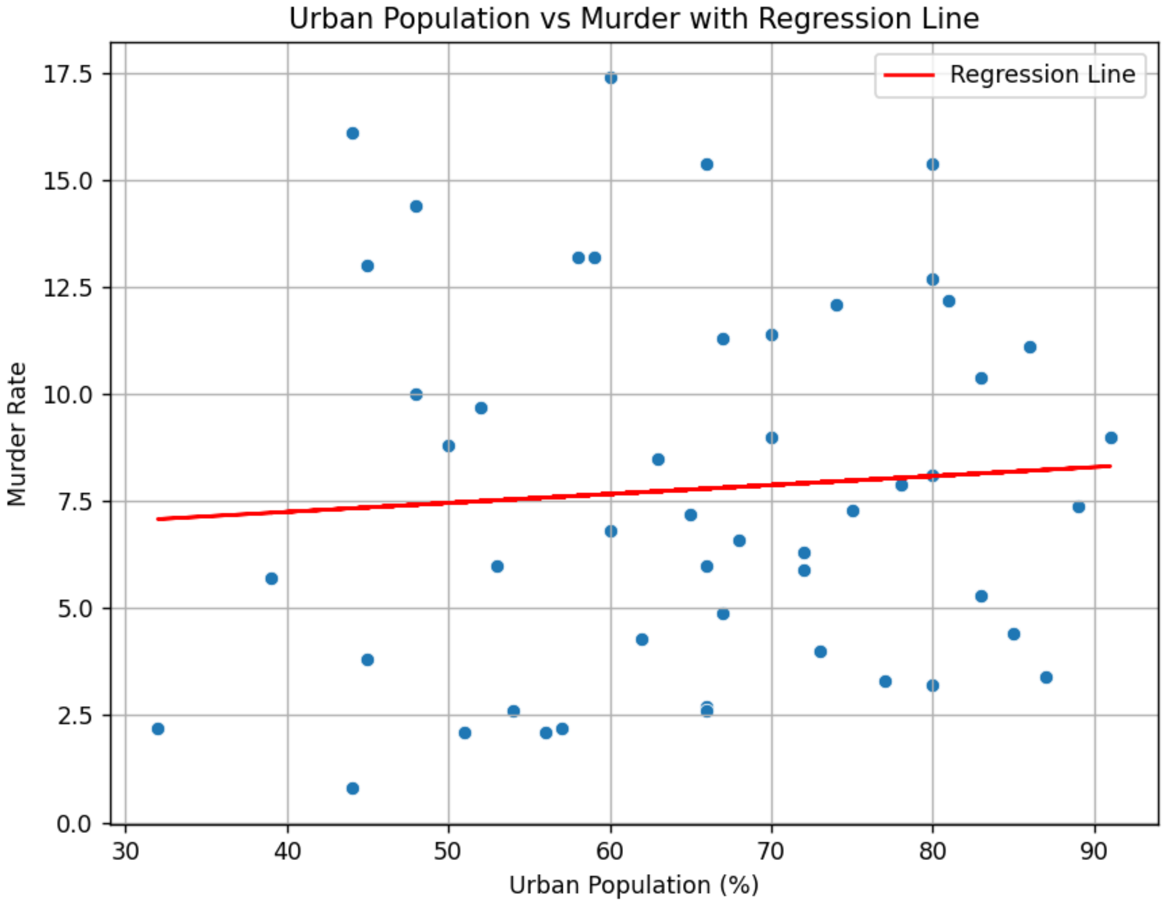
Scatterplots showing relationship of urban population vs each crime type with line of anomalies

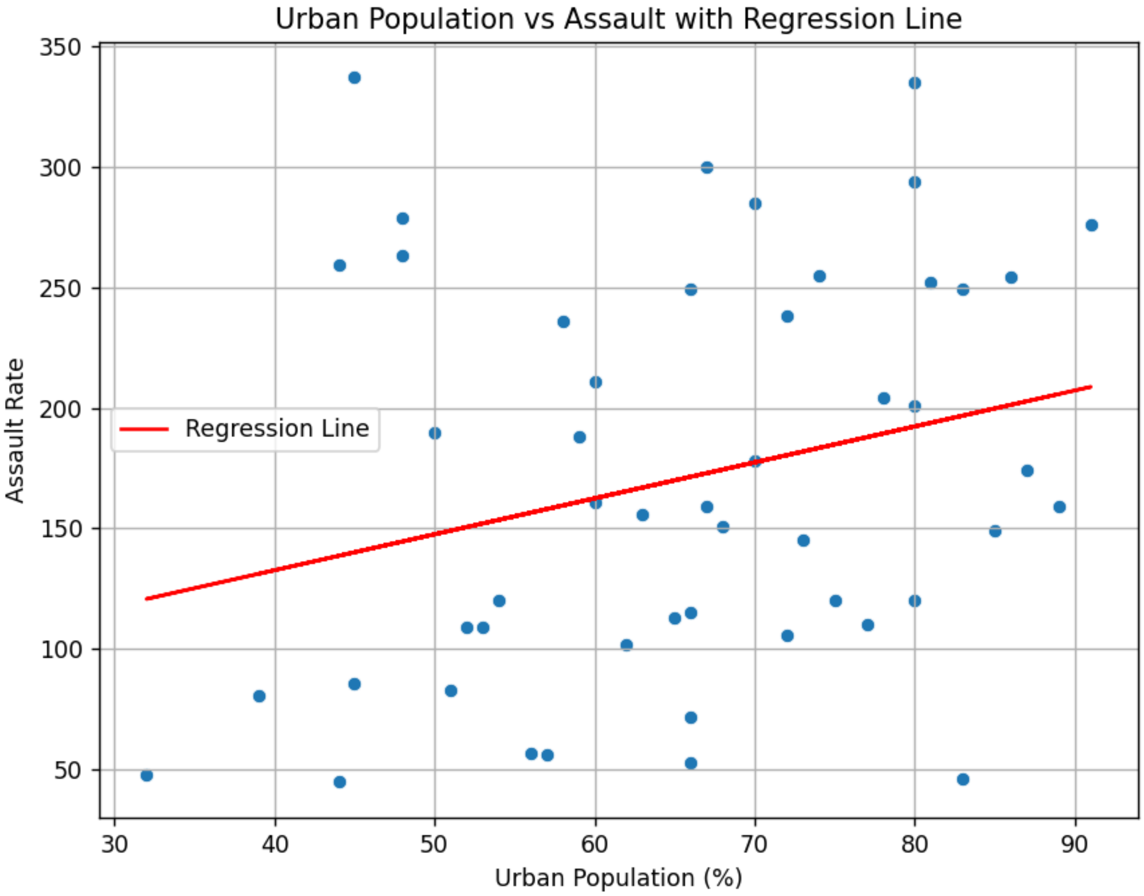


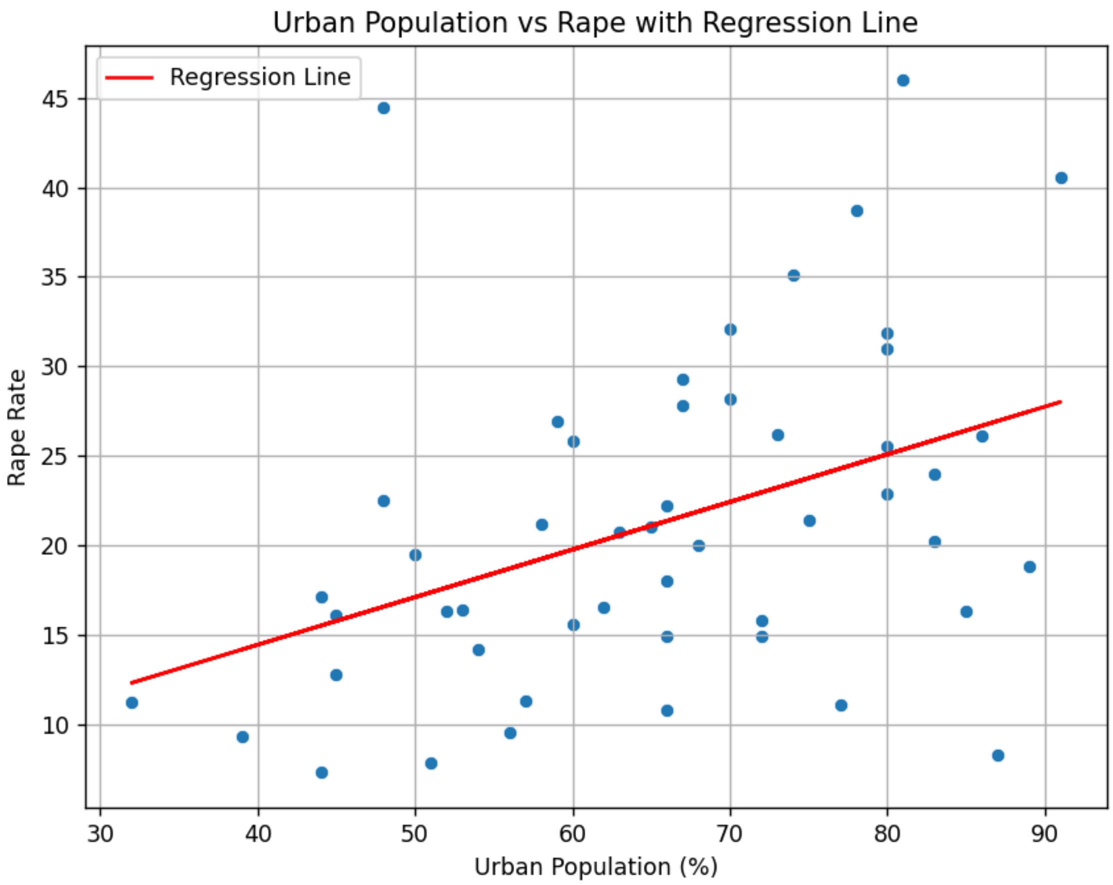

Urban Population vs Assault
Assault Rate
Urban Population (%)
50
100
150
200
250
300
350
30
40
50
60
70
80
90
North Carolina
Alaska
California
Nevada
Georgia
Vermont



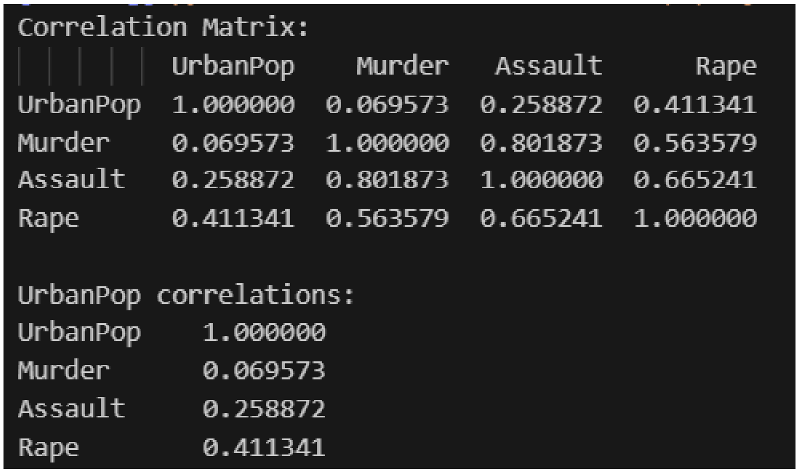
Scatterplots showing relationship of Urban population with the crime type with regression line


Urban Population vs Murder with Regression Line
Regression Line
Murder Rate
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
30
40
50
60
70
80
90
Urban Population (%)


Urban Population vs Assault with Regression Line
Assault Rate
Urban Population (%)
Regression Line
350
300
250
200
150
100
50
30
40
50
60
70
80
90



As above was the graphs below will show the calculations like the correlation matrix as well as the regression and even the anomalies



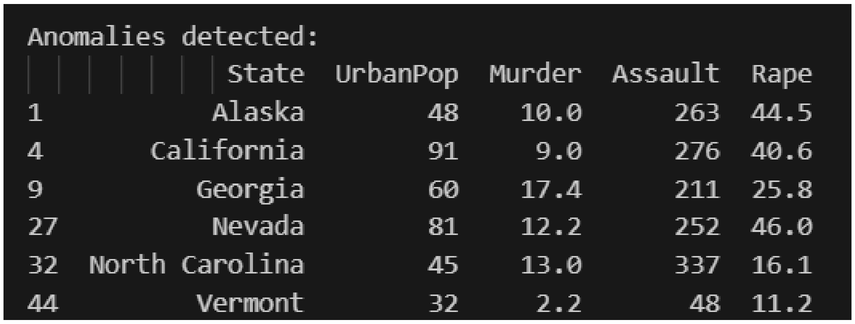
```
Correlation Matrix:
          UrbanPop    Murder   Assault      Rape
UrbanPop  1.000000  0.069573  0.258872  0.411341
Murder    0.069573  1.000000  0.801873  0.563579
Assault   0.258872  0.801873  1.000000  0.665241
Rape      0.411341  0.563579  0.665241  1.000000

UrbanPop correlations:
UrbanPop    1.000000
Murder      0.069573
Assault     0.258872
Rape        0.411341
```

if the value is closer 1 it indicates a strong positive correlation

if the value is closer to -1 then there is a strong negative correlation

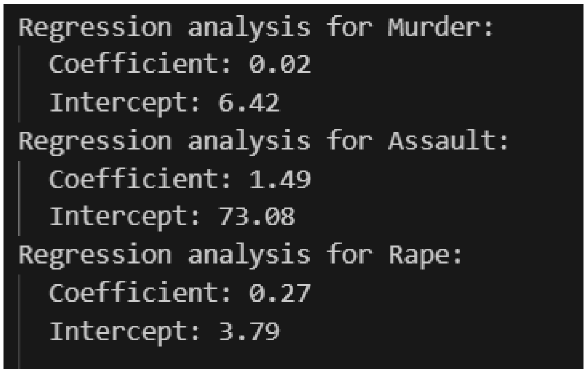
if the value is close to 0 then there is little to no linear relationship

Anomalies detected



```
Anomalies detected:
                   State  UrbanPop  Murder  Assault  Rape
1                 Alaska        48    10.0      263  44.5
4             California        91     9.0      276  40.6
9                Georgia        60    17.4      211  25.8
27                Nevada        81    12.2      252  46.0
32        North Carolina        45    13.0      337  16.1
44               Vermont        32     2.2       48  11.2
```

Regression analysis



```
Regression analysis for Murder:
  Coefficient: 0.02
  Intercept: 6.42
Regression analysis for Assault:
  Coefficient: 1.49
  Intercept: 73.08
Regression analysis for Rape:
  Coefficient: 0.27
  Intercept: 3.79
```

Discussion of how results:

First, by understanding the individualized crime by state is good to understand state wise how the distribution of crime is represented. Then by showing the relationship between urban population with crime shows how the densely populated areas have their own crime rates and show the distribution of crime. When looking at the correlation between urban population with crimes it shows rape only had the highest correlation with urban population while murder and assault really did not show as much of a correlation with urban population. When broken down for their own correlations of crime-to-crime murder and assault had their highest correlations between then because they had stronger positive correlations with each other. With anomalies of this project being Alaska, California, Georgia, Nevada, North Carolina, and Vermont. The regression analysis can be summed up to the idea that for the coefficient for every increase of 1% of urban population the crime rate increases by said coefficient while the intercept shows the idea of if there was 0% urban population how much of that crime would be present to show the inflation of the crime there really is in the US.

How this relates to the Project goals:

These results does show a relationship to the goals as listed before not only as the finding of these metrics were a goal but to show the idea if more policies are needed to stop these crimes because before the common misconception was always that urban population was the main cause of these crimes and that the crimes are heavily effected by the crimes while my findings actually find the opposite. Especially with the correlation matrix as there was only 1 crime with stronger positive relationship to urban population while others were not as correlated with urban population. The common misconception should be changed to other metrics like crime proportionality in the sense of if the states were able to crack down on assault they would effectively be able to crack down heavily on murder as well.

7.Discussion and Conclusions:

Project Issues

- Data constraints:
  - The dataset lacked additional variables such as income levels, unemployment rates, or education statistics, which could better explain variations in crime rates. This limited the depth of analysis.
  - Absence of temporal data restricted the ability to study trends over time, focusing the analysis solely on a snapshot.
- Linear Relationship Assumption
  - The regression analysis assumed a linear relationship between urban population and crime rates. While this approach provides initial insights, it may oversimplify relationships that are non-linear or influenced by other hidden factors.
- Outlier Influence
  - Certain states with extreme crime rates skewed regression results and visualizations. Although anomalies were flagged, their presence may impact on the overall conclusions.
- Geographic Data
  - The lack of latitude and longitude data prevented geospatial analysis, such as mapping crime clusters or identifying geographic patterns.

Limitations

- Correlation vs. Causation
  - The project reveals associations between urban population and crime rates but cannot establish causation. For example, higher urban populations might correlate with increased assault rates, but other factors (e.g., population density, socioeconomic disparity) might be the actual drivers.
- Simplified Crime Metrics
  - Aggregating crimes like Assault or Rape into single categories does not account for differences in severity or underlying causes within those categories.
- Scalability
  - The methods used work well for small datasets but may require optimization to handle larger or multi-year datasets.

Conclusion:

The project successfully analyzed the relationship between urban population and crime rates across US states, using visualizations, regression analysis, and anomaly detection. Results showed that rape rates have a strong positive correlation with urbanization, while murder and assault rates have weaker associations. Anomalies, such as states with high urban populations but unexpectedly low crime rates, provided unique insights into regional differences. The project highlighted the importance of data-driven decision-making for policymakers and demonstrated the utility of statistical and machine learning techniques learned during the course. Despite limitations like missing socioeconomic factors and the assumption of linearity, the project establishes a foundation for deeper, more comprehensive analyses in the future.