



TASK

Exploratory Data Analysis on the UsArrests Dataset

Visit our website

Contents

| | |
|--|-----------|
| Introduction | 3 |
| Data Cleaning | 3 |
| Missing Data | 5 |
| Data Stories And Visualizations | 6 |
| Visualization 1: Murder Rate Distribution by Number of States | 6 |
| Visualization 2: Assault Rate Distribution by Number of States | 7 |
| Visualization 3: Rape Rate Distribution by Number of States | 8 |
| Visualization 4: Urban Population Distribution by Number of States | 9 |
| Visualization 5: Correlation Heatmap of Crime Rates and Urban Population | 10 |
| Visualization 6: Top and Bottom 5 States by Crime and Urban Population Metrics .. | 12 |
| Visualization 7: Top 5 Most Dangerous vs. Safest States Based on Crime Index | 13 |
| Interactive Maps: Crime Rate Comparison Across States | 15 |

Introduction

The **UsArrests** (1973) dataset provides crime-related statistics for different states (or cities) in the United States, focusing on murder, assault, urban population percentage, and rape rates and has the following characteristics:

- **Number of rows:** 50
- **Number of columns:** 5
- **Column names:** City, Murder, Assault, UrbanPop, Rape
- **Data types:**
 - Categorical:**
 - City: Object (string) - Represents the name of the state.
 - Numerical:**
 - Murder: Float64 - The murder arrest rate (per 100,000 residents).
 - Assault: Int64 - The assault arrest rate (per 100,000 residents).
 - UrbanPop: Int64 - The percentage of the population living in urban areas.
 - Rape: Float64 - The rape arrest rate (per 100,000 residents).
- **Missing values:** There are no missing values in any column.
- **Sample Data:**
 - **City:** Nebraska
 - **Murder:** 4.3
 - **Assault:** 102
 - **UrbanPop:** 62%
 - **Rape:** 16.5

Data Cleaning

Data Cleaning Methods:

1. **Loading the Data:**
 - The dataset was loaded using `pd.read_csv()` and checked for any issues such as missing or duplicate rows.
2. **Duplicate Rows:**
 - The dataset was checked for duplicates using `us_arrests_df.duplicated()`. No duplicates were found.
3. **Missing Data:**
 - The dataset was scanned for missing values using `us_arrests_df.isna().any()`. No missing values were found.
4. **Renaming Columns:**
 - The column City was renamed to State for clarity using `us_arrests_df.rename()`.
5. **New Columns:**

- A new column `total_crime` was added, calculated by summing the values of the Murder, Assault, and Rape columns.
- Another column, `state_abbrev`, was created as a duplicate of the State column to assist with visualizations.

Visualizations Performed:

1. Grouped Bar Chart:

- A grouped bar chart was created to compare crime rates (Murder, Assault, Rape) for the top 5 most dangerous and safest states. Different colours were used to differentiate crime types, and bars were stacked for each state.

2. Choropleth Maps:

- Several choropleth maps were created using Plotly to visualize the geographical distribution of various metrics across states:
 - Urban population percentage (UrbanPop)
 - Murder rate (Murder)
 - Assault rate (Assault)
 - Rape rate (Rape)
 - Total crime rate (`total_crime`)
- Each map used a different colour scale to make the visual distinction clear.

3. Boxplots:

- Boxplots were used to visualize the distribution and detect outliers for each numerical column (Murder, Assault, UrbanPop, and Rape). These plots help identify states with extreme crime rates and population distributions.

4. Histograms:

- Histograms were created for the numerical columns to show the frequency distribution of crime rates and urban population percentages. These histograms provided insights into the central tendencies and spread of the data.

5. Heatmaps:

- A heatmap was generated to display the correlation matrix of the numerical columns (Murder, Assault, UrbanPop, and Rape). The heatmap used colour intensity to represent the strength of correlations, revealing relationships between the crime types and urban population.

6. Stacked Bar Chart:

- A stacked bar chart was created to show the contribution of each crime type (Murder, Assault, and Rape) to the total crime rate across different states. The bars were stacked to give a clear picture of the overall crime composition per state, with different colours representing each crime type.

Missing Data

The **UsArrests** dataset was checked for missing data during the EDA process, and the following steps were taken:

Checking for Missing Data:

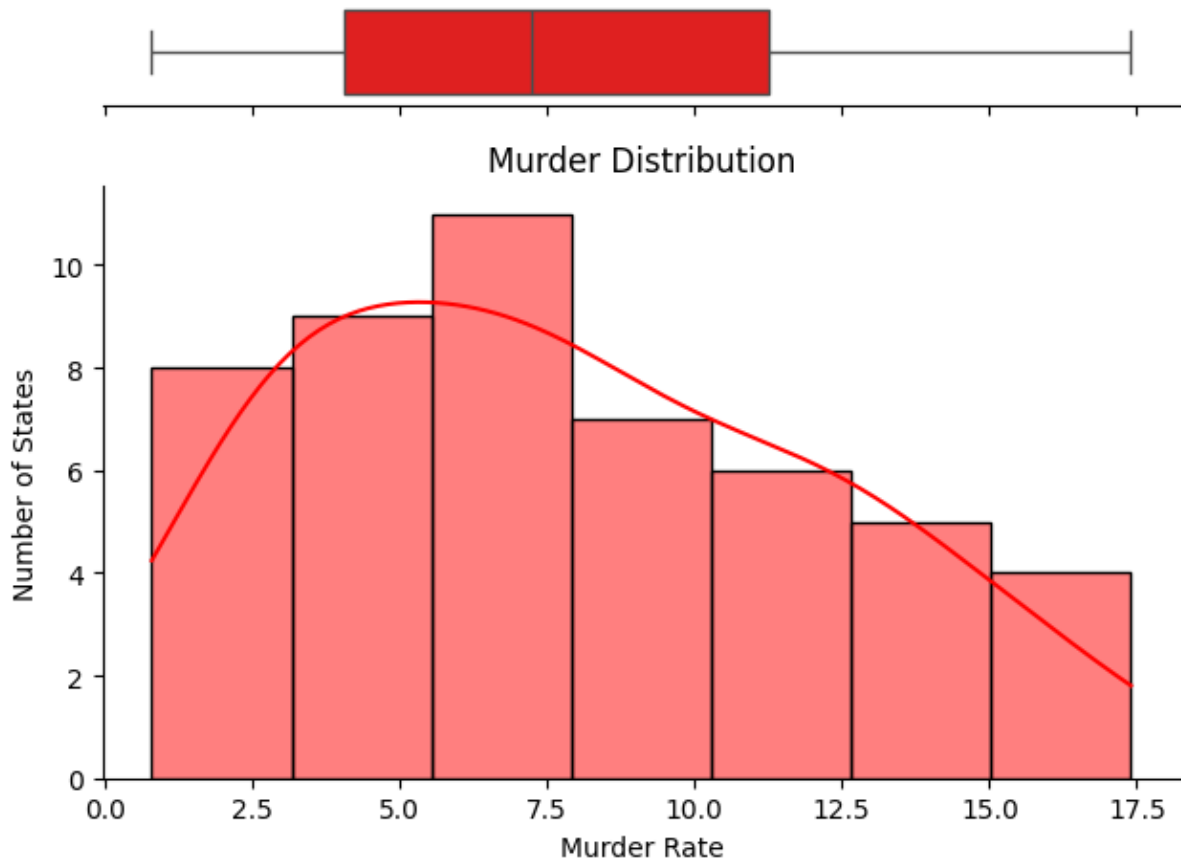
- The dataset was scanned using `us_arrests_df.isna().any(axis=1)` to identify any rows that had missing values.
- After the check, it was confirmed that there were **no missing values** in any column.

How Missing Data Was Handled:

Since there were no missing values in the dataset, no specific missing data handling techniques (such as imputation or deletion) were necessary. The data was complete, and no further actions were required in this aspect.

Data Stories And Visualizations

Visualization 1: Murder Rate Distribution by Number of States



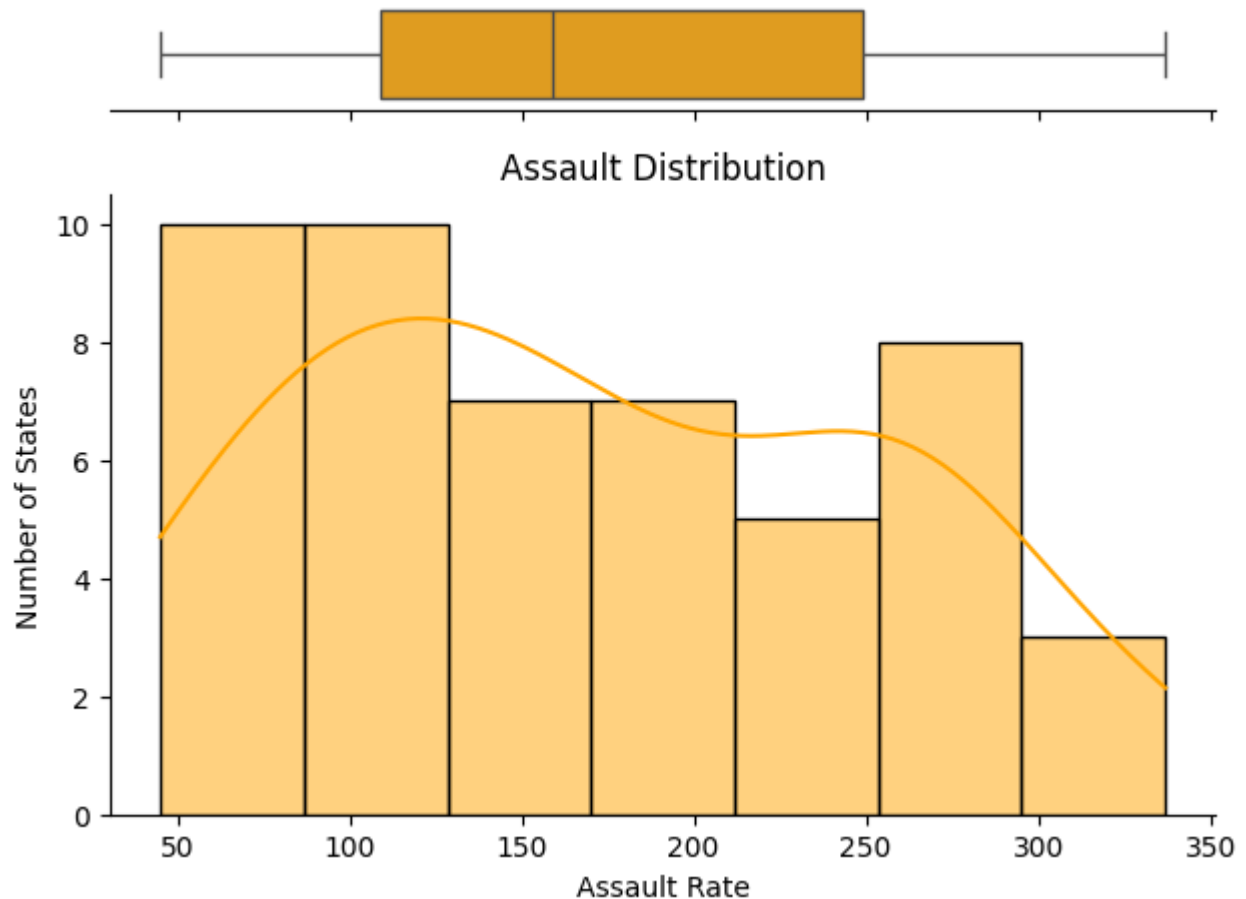
This visualization shows the distribution of murder rates across different U.S. states.

- The top box plot indicates that the middle 50% of states have murder rates between about 3.5 and 9.5, with a median around 7.
- The histogram beneath it reveals a right-skewed distribution, with most states having murder rates between 2.5 and 10. Fewer states experience higher murder rates (above 10).
- The density curve further emphasizes this skewness, showing a peak around 6-7.

Assumptions:

- Most states tend to have relatively moderate murder rates.
- There are fewer outliers with higher murder rates, as indicated by the right tail.
- The data is not normally distributed but right-skewed.

Visualization 2: Assault Rate Distribution by Number of States



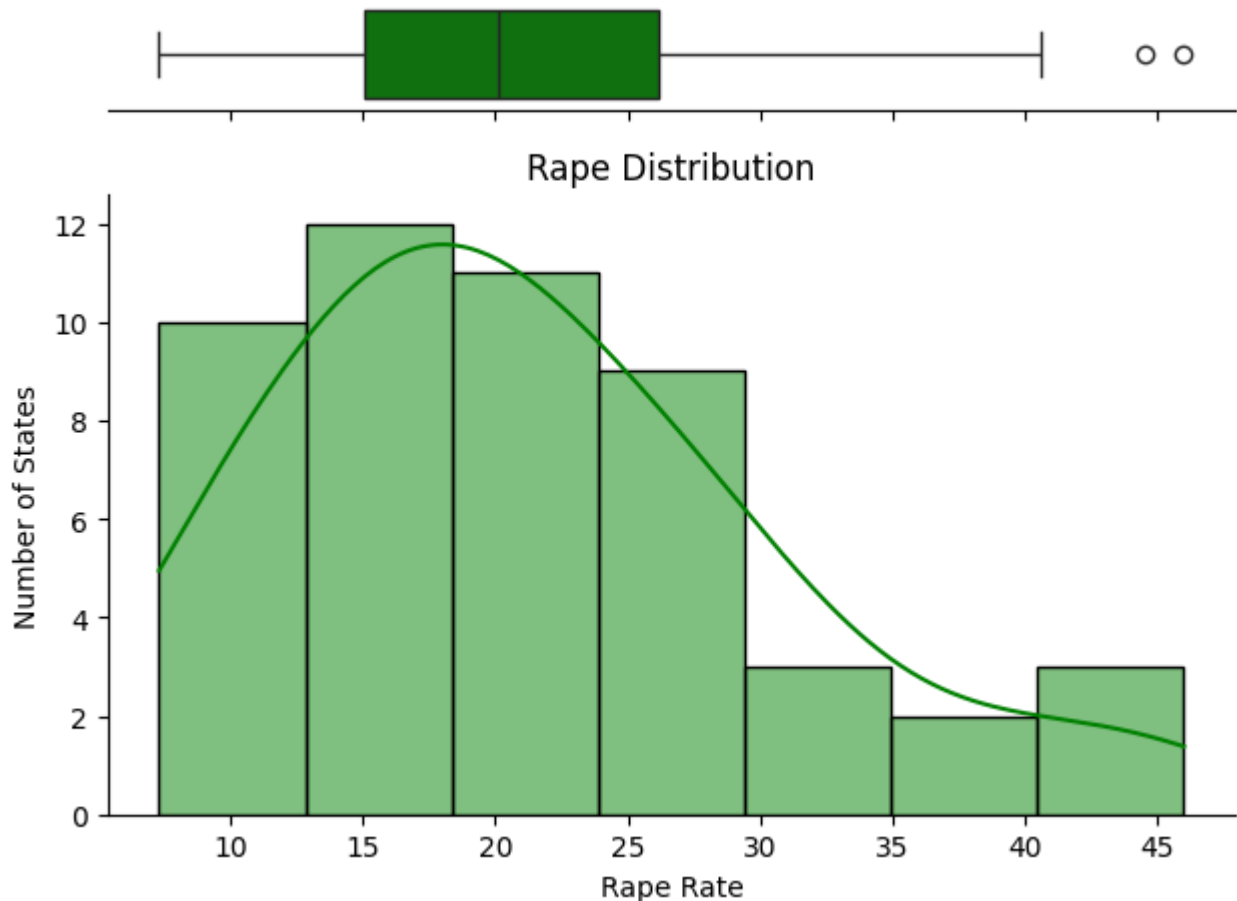
This visualization shows the distribution of assault rates across different U.S. states:

- The box plot at the top shows that the middle 50% of states have assault rates between approximately 100 and 250, with a median near 150.
- The histogram shows a bimodal distribution, with peaks around 90 and 260, indicating two groups of states: one with lower assault rates (50-150) and another with higher rates (200-300).
- The density curve confirms these two peaks and suggests that fewer states experience extreme assault rates, especially beyond 300.

Assumptions:

- There may be two distinct clusters of states regarding assault rates.
- Assault rates are more evenly distributed compared to murder rates, but some states still have notably high rates.
- The distribution is somewhat spread out.

Visualization 3: Rape Rate Distribution by Number of States



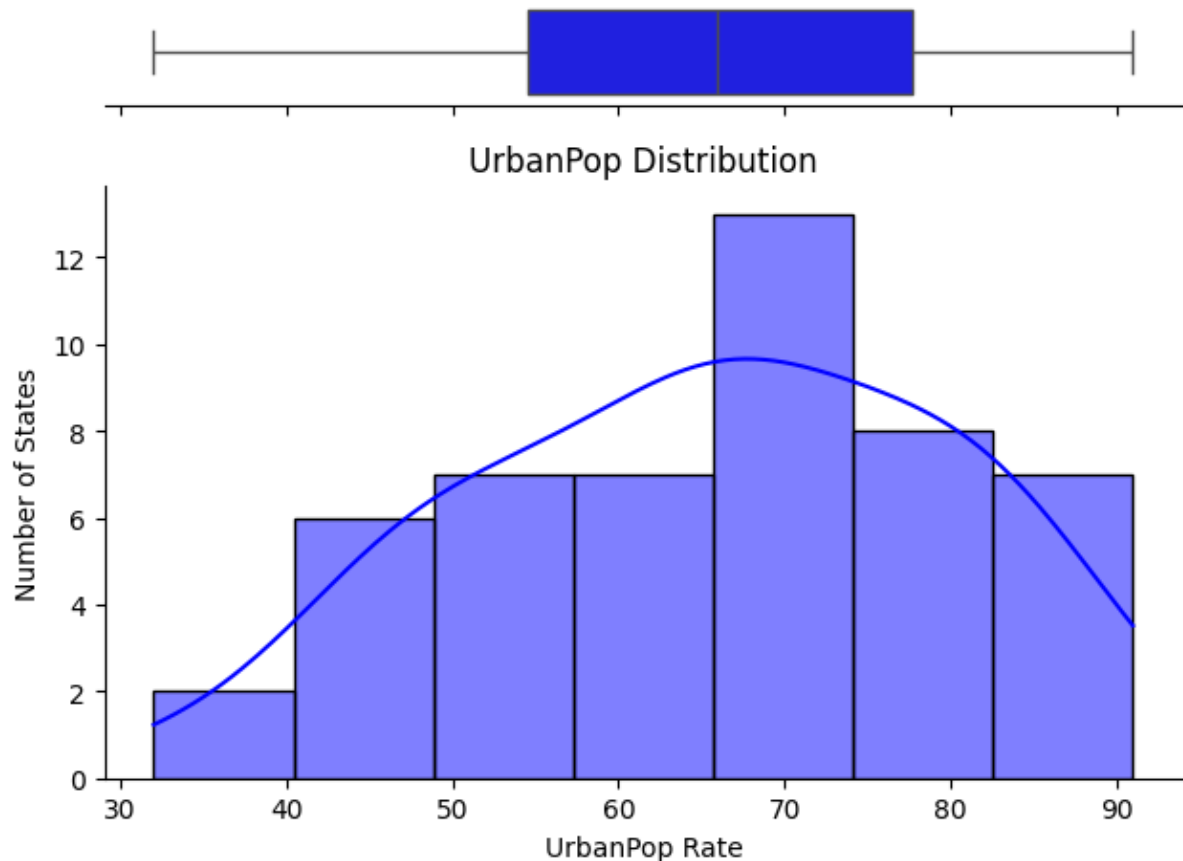
This visualization shows the distribution of rape rates across different U.S. states:

- The box plot indicates that the middle 50% of states have rape rates between approximately 15 and 27, with a median around 20. There are two outliers on the high end, as shown by the points to the right.
- The histogram shows a left-skewed distribution, with most states having rape rates between 10 and 25. The frequency drops significantly for rates above 30.
- The density curve aligns with this, showing a peak around 17-20 and a sharp decline as rates increase.

Assumptions:

- Most states have moderate rape rates, with few states showing extremely high rates, evidenced by the outliers.
- The data is slightly left-skewed, suggesting that higher rape rates are rare but present in a few outlying states.

Visualization 4: Urban Population Distribution by Number of States



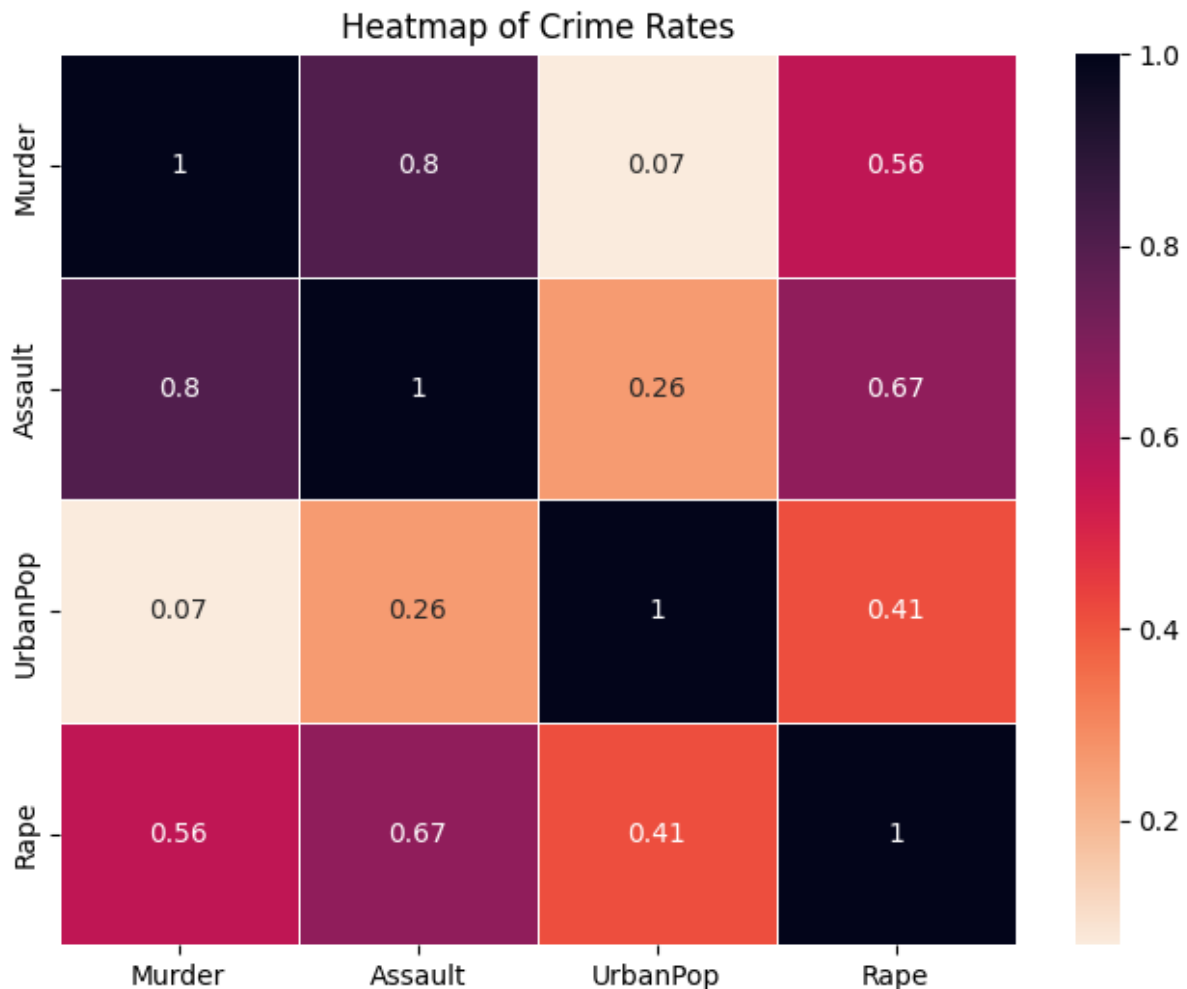
This visualization shows the distribution of urban population (UrbanPop) rates across different U.S. states:

- The box plot at the top shows that the middle 50% of states have urban population rates between about 55 and 80, with a median around 70.
- The histogram shows a fair distribution, with most states having urban population rates between 50 and 80. There is a peak around 70.
- The density curve suggests a slight right-skew, but overall, the distribution is close to normal.

Assumptions:

- Most states have a moderately high level of urbanization, with few outliers.
- Urban population rates are fairly evenly distributed, with no extreme skew or significant clustering at the lower or higher ends of the distribution.

Visualization 5: Correlation Heatmap of Crime Rates and Urban Population



This heatmap shows the correlation between various crime rates (**Murder**, **Assault**, **Rape**) and the urban population percentage (**UrbanPop**) across U.S. states. Correlation values range from -1 to 1, where 1 indicates a perfect positive correlation and -1 indicates a perfect negative correlation. Values closer to 0 imply no correlation.

- **Murder and Assault (0.8 correlation):**
 - There is a strong positive correlation between **Murder** and **Assault**, suggesting that states with high murder rates also tend to have high assault rates. This could indicate common underlying socio-economic or environmental factors that drive both forms of violent crime simultaneously. These factors might include poverty, social instability, or ineffective law enforcement in certain regions.
- **Rape's Relationship to Other Crimes:**
 - **Rape** shows a moderate positive correlation with both **Assault** (0.67) and **Murder** (0.56). This implies that states with higher rates of assault and murder are more likely to also have higher rates of rape. While these

are distinct crimes, the correlation suggests that some states may struggle with multiple forms of violent crime, driven by similar underlying factors like poor public safety infrastructure, social issues, or regional violence trends.

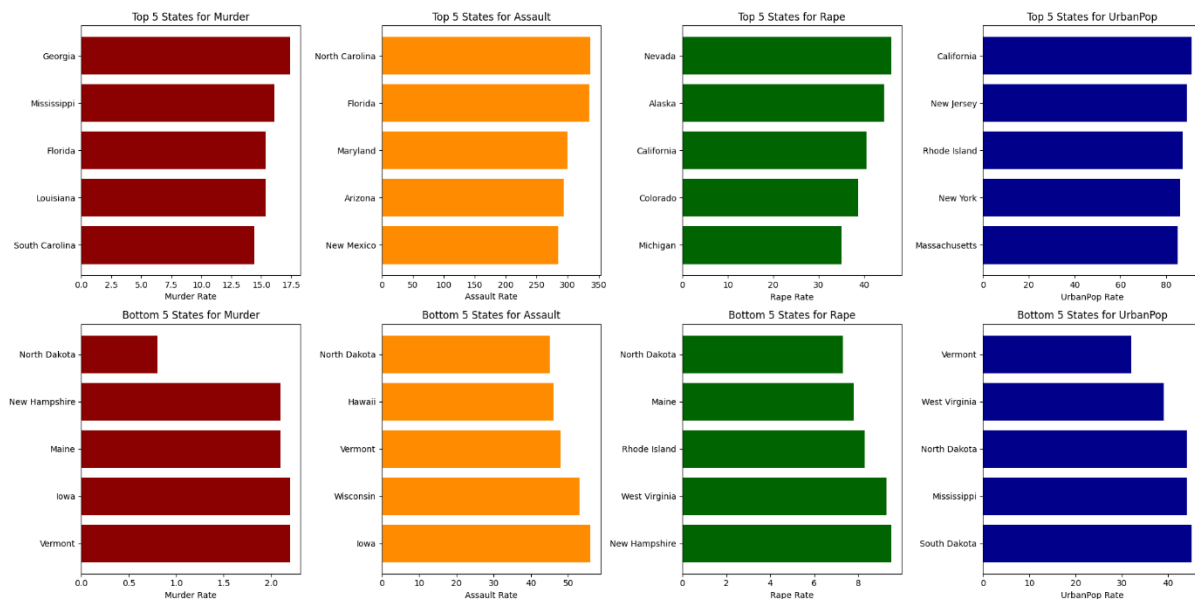
- **Urban Population (UrbanPop) and Crime:**

- Interestingly, **UrbanPop** shows low correlations with most crime rates. It is least correlated with **Murder** (0.07) and **Assault** (0.26), indicating that a state's level of urbanization does not have a significant effect on these violent crimes.
- The correlation between **UrbanPop** and **Rape** (0.41) is moderate, suggesting that urbanized states might have slightly higher rates of rape, though this correlation is still not particularly strong. This could be due to differences in population density, reporting practices, or policing in urban areas compared to rural ones.

Assumptions:

- **Clustering of Violent Crimes:** The strong correlation between Murder, Assault, and Rape suggests that states with a higher propensity for one violent crime often face challenges with other violent crimes as well. These crimes may share common drivers such as socio-economic inequality, lack of access to education, or strained relationships between communities and law enforcement.
- **Urbanization's Weak Influence:** The low correlation between **UrbanPop** and most crime rates implies that urbanization itself is not a significant predictor of violent crime rates. While cities may have more incidents of crime simply due to higher populations, the per capita crime rates are not strongly linked to how urbanized a state is. This could indicate that factors such as law enforcement effectiveness, social programs, and community support play larger roles in managing crime in both urban and rural areas.

Visualization 6: Top and Bottom 5 States by Crime and Urban Population Metrics



This visualization highlights the top and bottom 5 states for murder, assault, rape, and urban population rates:

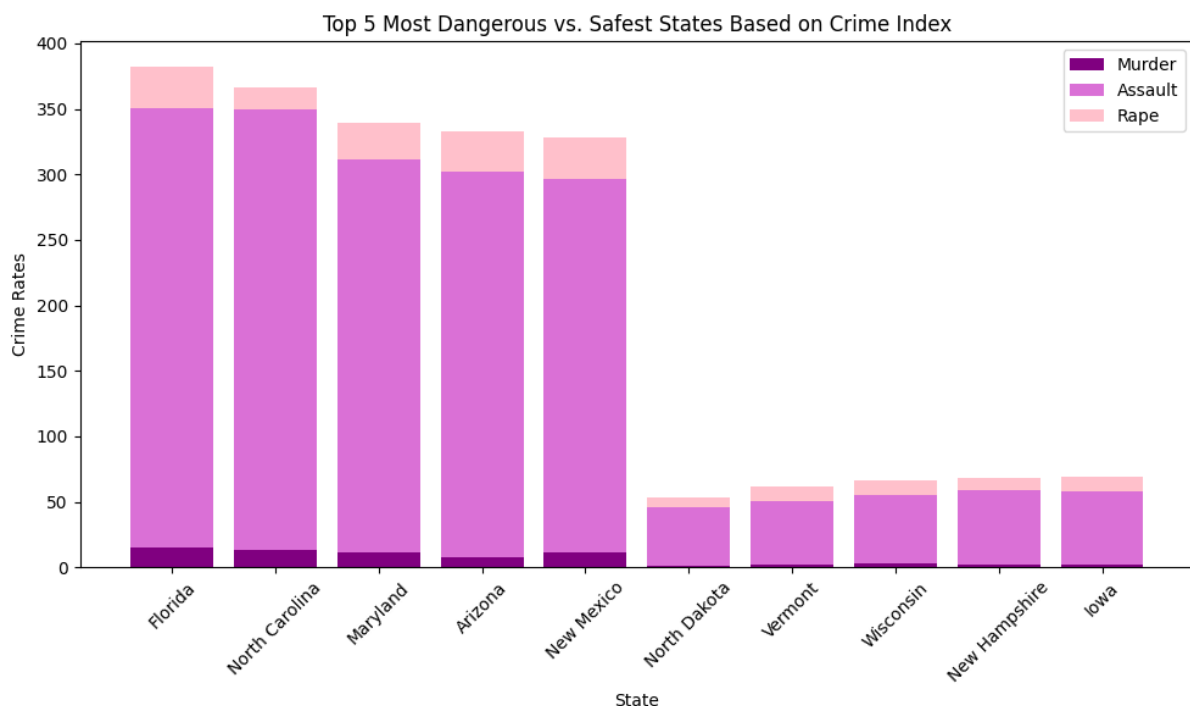
- **Top 5 States for Murder:** Southern states like Georgia, Mississippi, and South Carolina have the highest murder rates, suggesting that regional factors, such as economic disparities or cultural dynamics, may contribute to violent crime.
- **Top 5 States for Assault:** States such as North Carolina and Florida rank highest in assault rates, indicating that certain regions may experience more violent crime, possibly due to social or environmental stressors.
- **Top 5 States for Rape:** Nevada and Alaska report the highest rape rates, potentially reflecting specific issues related to crime or the way incidents are reported in these states.
- **Top 5 States for Urban Population:** Highly urbanized states like California, New Jersey, and New York dominate this category, reflecting their large urban centres and population density.
- **Bottom 5 States for Murder, Assault, and Rape:** States like North Dakota and Vermont consistently appear in the bottom 5, indicating lower rates of violent crime in these areas.
- **Bottom 5 States for Urban Population:** Rural states like Vermont, West Virginia, and North Dakota have the lowest urbanization rates, which often correlates with a lower incidence of certain crimes.

Assumptions:

- Crime rates may be influenced by a combination of regional and demographic factors, with southern and highly urbanized states tending to report higher rates of violent crime.

- Lower urban populations often coincide with lower crime rates, although urbanization itself does not necessarily lead to more crime. Other factors, such as local economies and social policies, also play a role.
- It's important to note that crime reporting and arrest rates can affect these statistics. States with higher arrest rates may not necessarily be more dangerous; they may have better crime reporting systems and higher rates of law enforcement intervention.
- Biases and discrimination can impact arrest rates and might not fully reflect the actual safety or crime levels in a state.

Visualization 7: Top 5 Most Dangerous vs. Safest States Based on Crime Index



This bar chart visualizes the crime index for the top 5 most dangerous and the top 5 safest states, combining **murder**, **assault**, and **rape** rates to present a comprehensive picture of violent crime in these regions.

Most Dangerous States

The states identified as the most dangerous - **Florida**, **North Carolina**, **Maryland**, **Arizona**, and **New Mexico** - stand out due to their significantly higher overall crime rates. These rates are driven predominantly by:

- **Assault:** Assault is the largest contributor to crime in these states, especially in **Florida**, which has the highest overall crime rate in the chart. This pattern

suggests that violent confrontations and physical attacks are major concerns in these regions.

- **Rape:** States like **Florida** and **North Carolina** also see high rape rates, adding to their crime index.
- **Murder:** Though murder contributes less than the other categories in terms of raw numbers, it remains a critical factor, especially in states with a high rate of all three violent crimes. States like **Arizona** show lower murder rates compared to their assault rates, but the overall contribution of all violent crimes keeps their crime index high.

Safest States

On the other hand, the states identified as the safest - **North Dakota**, **Vermont**, **Wisconsin**, **New Hampshire**, and **Iowa** - show considerably lower crime rates across all categories. These states are characterized by:

- **Low Assault Rates:** Assault remains the dominant crime in the safest states, but the total number of incidents is significantly lower compared to the most dangerous states.
- **Minimal Murder and Rape Rates:** Murder and rape rates in these states are minimal, contributing to their position as low-risk regions. For instance, **North Dakota** and **Vermont** exhibit very low murder rates, reinforcing their ranking among the safest states.

Assumptions:

- **Assault Drives Crime Index:** The chart suggests that **assault rates** are the key factor differentiating the most dangerous states from the safest ones. States with higher assault rates, such as **Florida** and **North Carolina**, tend to top the list of dangerous states, while states with lower assault rates tend to rank as safer.
- **Combination of Crimes:** While assault is the most prevalent crime, the combination of all three categories - murder, assault, and rape - determines the overall crime index. The most dangerous states often have higher rates across the board, indicating a broader issue with violent crime.
- **Regional and Demographic Factors:** The most dangerous states are largely found in the **southern and western regions**, suggesting that regional factors (such as economic conditions, population density, or law enforcement policies) may contribute to the higher crime rates. In contrast, the safest states are generally more rural, with **lower population density**, which may play a role in their reduced crime rates.
- **Social Implications:** The lower crime rates in states like **Vermont** and **Iowa** suggest that social stability, economic conditions, and community cohesion might help mitigate violent crime. Conversely, states with higher crime rates

may require more focused interventions, including crime prevention programs, community engagement, and targeted law enforcement strategies.

Interactive Maps: Crime Rate Comparison Across States

Crime Rate Comparison Across States:

Which states experience the highest and lowest crime rates? Are specific crimes more concentrated in certain areas?

These interactive choropleth maps provide an in-depth exploration of the dataset. By hovering over each state, you can view detailed information, including exact crime rates. The colour intensity in the heatmaps reflects the severity of the crime rate - darker colours indicate higher rates.

To explore each crime category in more detail, click on the links below for the corresponding choropleth maps:

- [Urban Population Distribution](#)
- [Murder Rates by State](#)
- [Assault Rates by State](#)
- [Rape Rates by State](#)
- [Total Crime Rates by State](#)

These maps allow for a comprehensive comparison of crime patterns across states, revealing geographic trends and potential correlations between crime and population density.

Alternatively, you can download and open the map from the attached html file.

THIS REPORT WAS WRITTEN BY : Christen Rao
