**Data Cleaning Workflow**

1. Importing Libraries: The notebook begins by importing essential libraries such as Pandas, Matplotlib, and Seaborn.

2. Loading Data: The dataset crime\_data.csv is loaded into a Pandas DataFrame.

3.Displaying Data: It displays the first few rows of the dataset, which contains columns like 'Murder,' 'Assault,' 'UrbanPop,' and 'Rape,' associated with different locations (states).

4.Data Inspection:

* The df.info() function is used to inspect the structure of the data, ensuring all fields have the correct data types (e.g., float, int, object).
* The df.isnull().sum() function checks for any missing values in the dataset, revealing that there are no nulls.

5.Cleaning Data:

* The first column is renamed to "Location."
* The code checks for and confirms that there are no duplicates in the dataset.

6.Saving the Cleaned Data: After the data is cleaned, it is saved as Cleaned\_Dataset.csv.

**EDA Workflow:**

**1.Summary Statistics**

|  | **Murder** | **Assault** | **UrbanPop** | **Rape** |
| --- | --- | --- | --- | --- |
| **count** | **50.00000** | **50.000000** | **50.000000** | **50.000000** |
| **mean** | **7.78800** | **170.760000** | **65.540000** | **21.232000** |
| **std** | **4.35551** | **83.337661** | **14.474763** | **9.366385** |
| **min** | **0.80000** | **45.000000** | **32.000000** | **7.300000** |
| **25%** | **4.07500** | **109.000000** | **54.500000** | **15.075000** |
| **50%** | **7.25000** | **159.000000** | **66.000000** | **20.100000** |
| **75%** | **11.25000** | **249.000000** | **77.750000** | **26.175000** |
| **max** | **17.40000** | **337.000000** | **91.000000** | **46.000000** |

**2.Highest and lowest cities for respective crimes**

Crime: Assault

Highest in: North Carolina (337)

Lowest in: North Dakota (45)

Crime: Murder

Highest in: Georgia (17.4)

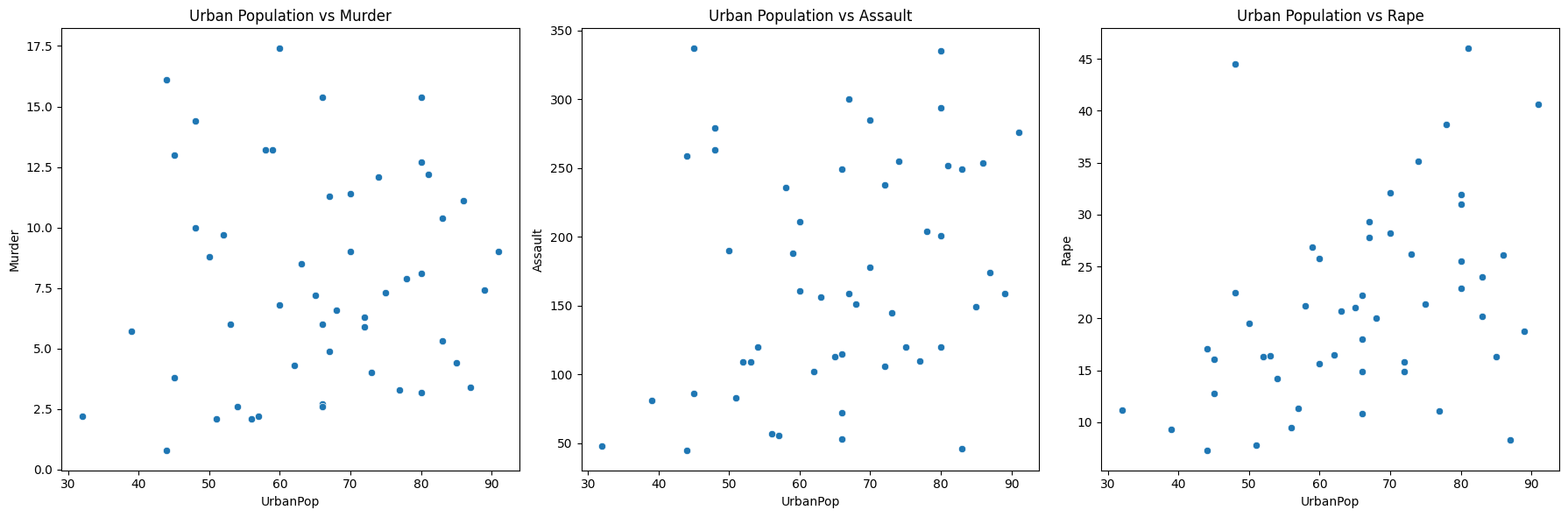
Lowest in: North Dakota (0.8)

Crime: Rape

Highest in: Nevada (46.0)

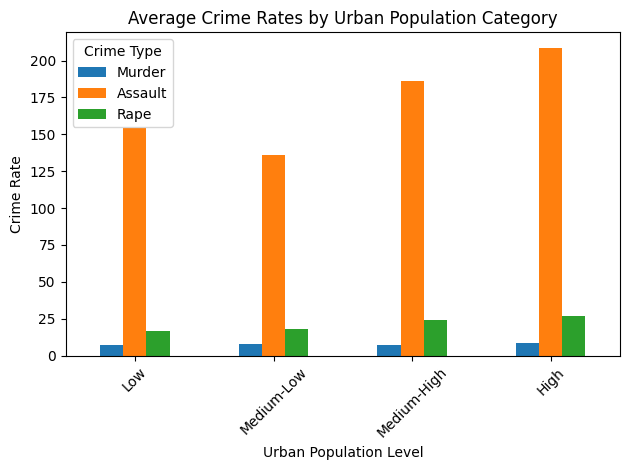
Lowest in: North Dakota (7.3)

**3.Urban Population vs crime**

****

**No specific relation between Urban population and Crime rate**

**4.Average Crime Rates by population category**

**Categrised Population into four categories and Assault has been the most frequently done crime**

**5.Average crime rates by urban population category:**

**Murder Assault Rape**

**Urban\_Category**

**Low 7.323077 154.538462 16.538462**

**Medium-Low 7.815385 136.076923 18.030769**

**Medium-High 7.463636 186.090909 23.809091**

**High 8.500000 208.692308 26.946154**

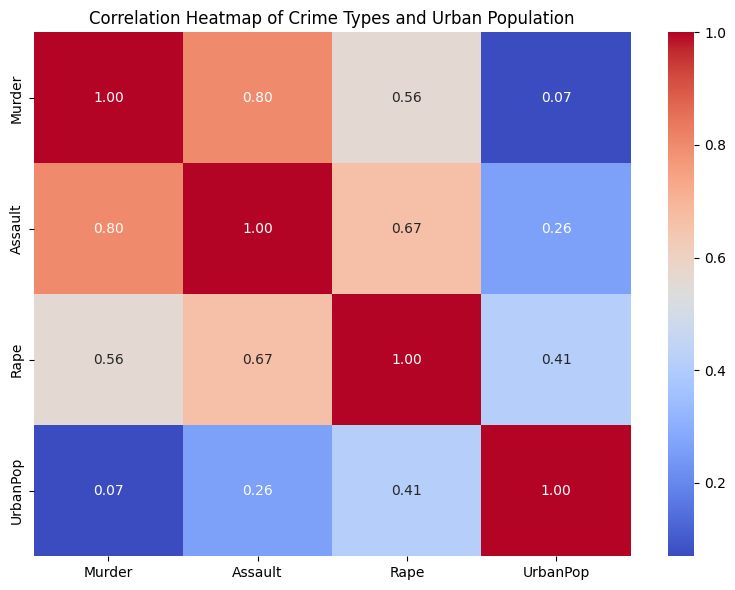
**Percentage increase from lowest to highest urban population:**

**Murder 16.07**

**Assault 35.04**

**Rape 62.93% increase**

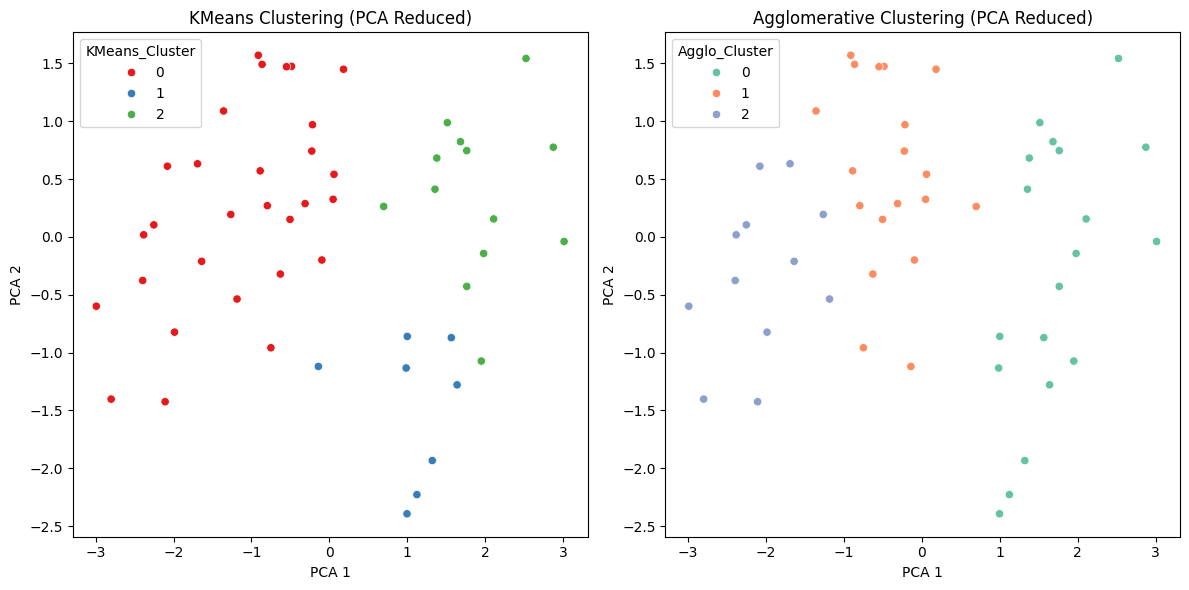
**6..HeatMap**

****

**Strong correlation between (Assault,Murder) and (Assault,Rape)**

**Advanced EDA and ML**

**1.Clustering Locations**

****

**Used Kmeans clustering for Clustering after Principal Component Analysis**

**2.Combined Crime Index:**

**Top 5 states by Crime Index:**

**Location Crime\_Index**

**8 Florida 3.321887**

**27 Nevada 3.226160**

**4 California 3.145537**

**21 Michigan 2.830112**

**30 New Mexico 2.745367**

**Bottom 5 states by Crime Index:**

**Location Crime\_Index**

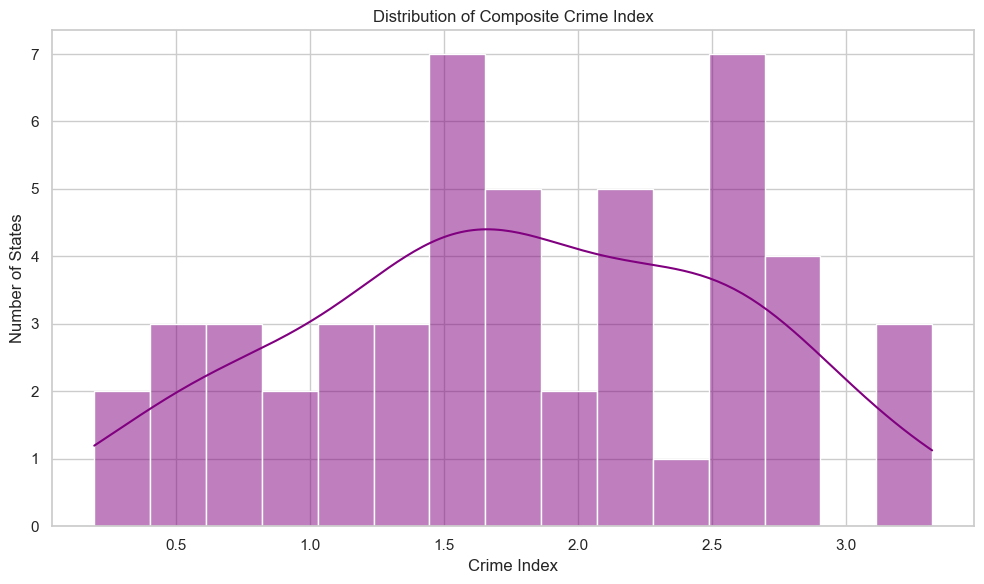
**44 Vermont 0.195387**

**33 North Dakota 0.203390**

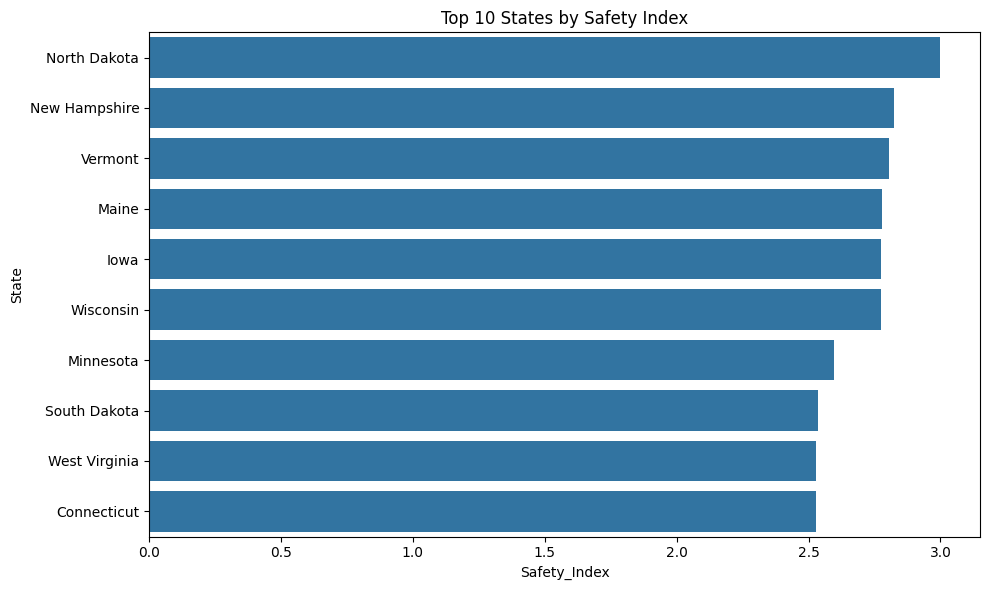
**18 Maine 0.543404**

**28 New Hampshire 0.583036**

**47 West Virginia 0.588792**

****

**3.Safety Index:**

****

**Top 5 Safest States:**

**State Safety\_Index**

**33 North Dakota 3.000000**

**28 New Hampshire 2.823743**

**44 Vermont 2.804613**

**18 Maine 2.778630**

**14 Iowa 2.774632**

**Bottom 5 States (Need Most Improvement):**

**State Safety\_Index**

**21 Michigan 0.881753**

**4 California 0.854463**

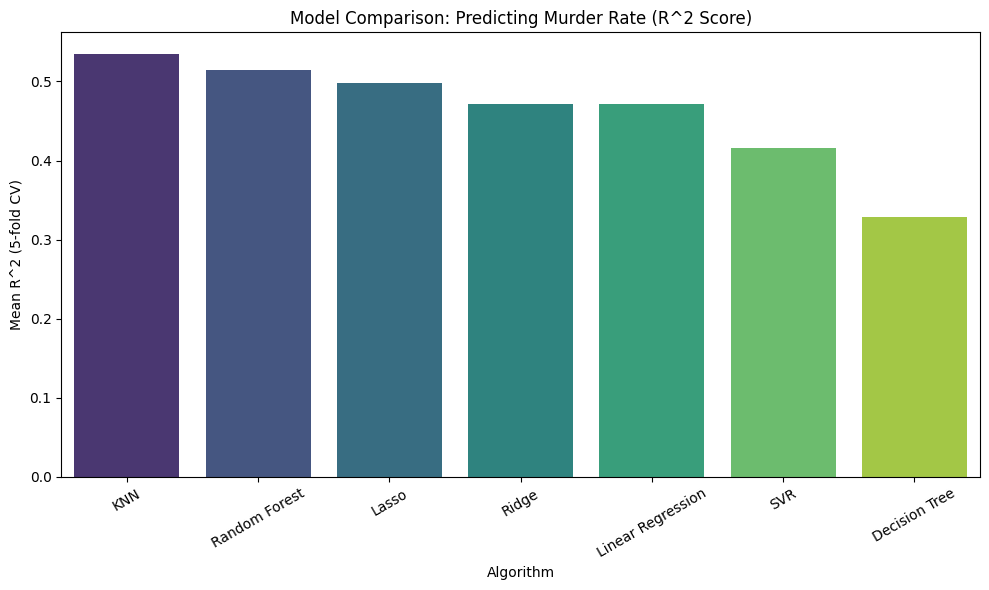
**1 Alaska 0.737967**

**27 Nevada 0.604349**

**8 Florida 0.491672**

**4.Model Building:**

**evaluated the performance of several regression models (Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, KNN, and SVR) in predicting the 'Murder' rate based on 'Assault', 'UrbanPop', and 'Rape' features from a DataFrame (df). It employs 5-fold cross-validation to obtain robust estimates of each model's R-squared score. The mean and standard deviation of these scores are calculated and stored. Finally, the code presents the results in a Pandas DataFrame sorted by the mean R-squared and visualizes the mean R-squared scores using a bar plot, identifying and printing the model with the highest mean R-squared as the best predictor.**

****

**Mean R2 Std R2**

**KNN 0.535295 0.233643**

**Random Forest 0.514954 0.242073**

**Lasso 0.498541 0.220690**

**Ridge 0.471922 0.255859**

**Linear Regression 0.471860 0.255910**

**SVR 0.415781 0.219290**

**Decision Tree 0.328097 0.226634**

**KNN has been the best performing model used hyerparameter tuning too for optimized parameters.**

**5.Regional Analysis**

**Murder Assault Rape CompositeCrimeIndex**

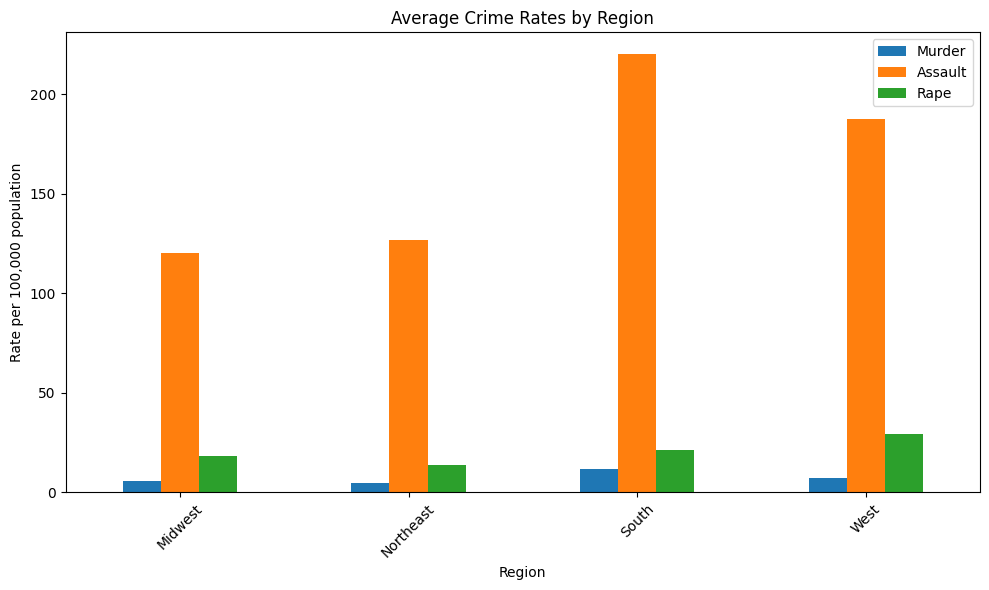
**Region**

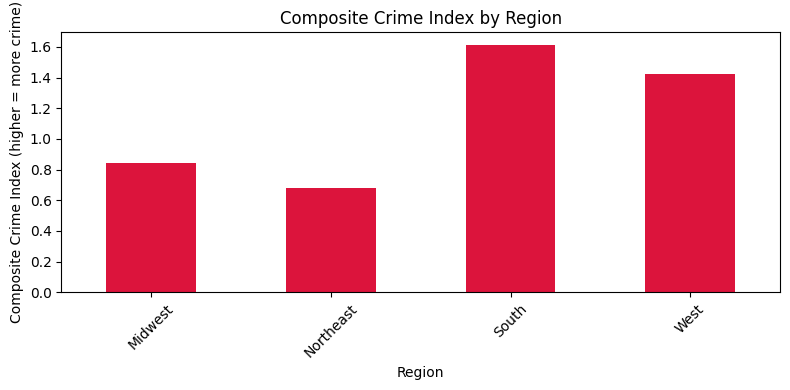
**Midwest 5.700000 120.333333 18.441667 0.841070**

**Northeast 4.700000 126.666667 13.777778 0.682005**

**South 11.706250 220.000000 21.162500 1.614522**

**West 7.030769 187.230769 29.053846 1.424554**

****

****

**Murder Assault Rape CompositeCrimeIndex**

**Region**

**Midwest 5.700000 120.333333 18.441667 0.841070**

**Northeast 4.700000 126.666667 13.777778 0.682005**

**South 11.706250 220.000000 21.162500 1.614522**

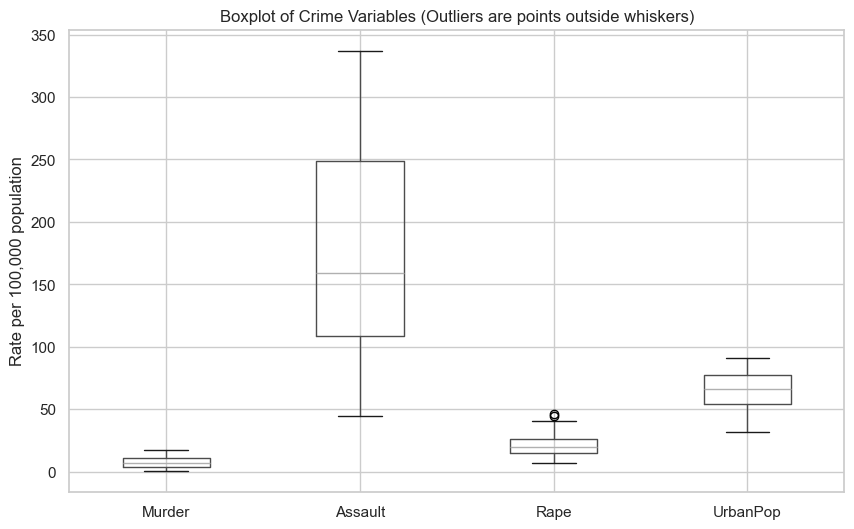
**West 7.030769 187.230769 29.053846 1.424554**

**South Region has been the most violent region in terms of crime**

**6.Outliers Detection:**

**a.Using Inter-Quartile Region Approach**

**The Interquartile Range (IQR) method is a robust, non-parametric technique for outlier detection that focuses on the spread of the central 50% of the data. It calculates the IQR as the difference between the third quartile (Q3) and the first quartile (Q1): IQR=Q3−Q1. Outliers are identified as data points falling below the lower bound (Q1−1.5×IQR) or above the upper bound (Q3+1.5×IQR). This approach effectively flags values that are significantly distant from the typical range of the data, making it particularly useful for datasets with non-normal distributions where standard deviation-based methods might be less reliable.**

****

**Outliers Detected:**

**Two outlier states has been detected for Rape type i.e; “ALASKA,NEVADA”**

**Q1 Q3 IQR Lower Bound Upper Bound Num Outliers**

**Murder 4.075 11.25 7.175 -6.6875 22.0125 0**

**Assault 109.0 249.0 140.0 -101.0 459.0 0**

**Rape 15.075 26.175 11.1 -1.575 42.825 2**

**UrbanPop 54.5 77.75 23.25 19.625 112.625 0**

**Outlier states for Murder : []**

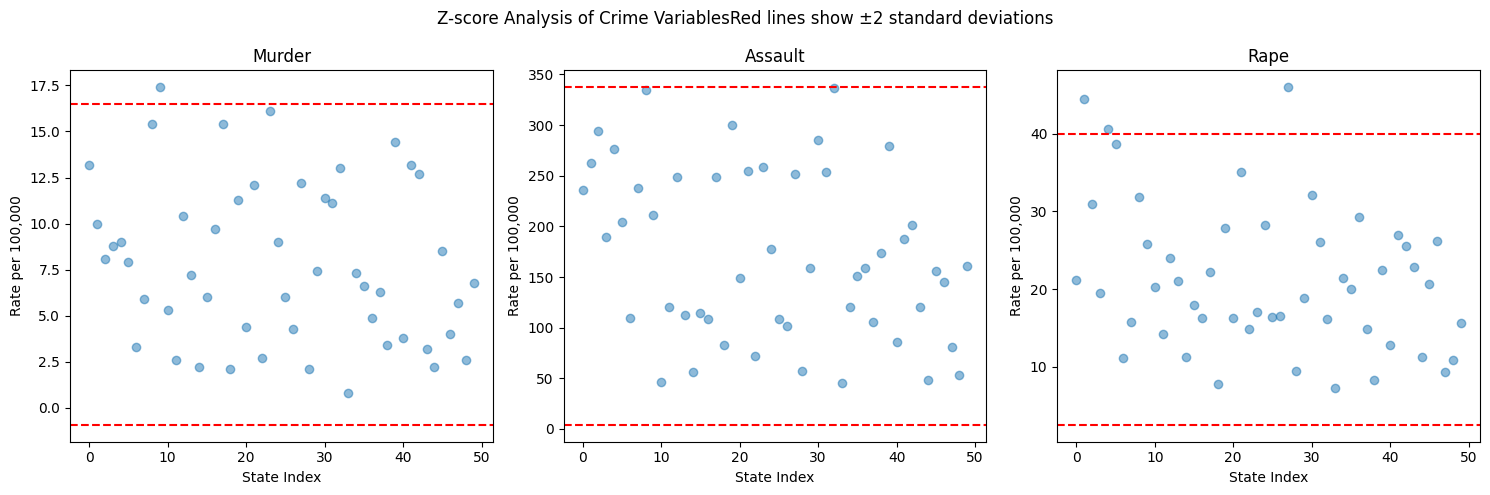
**Outlier states for Assault : []**

**Outlier states for Rape : ['Alaska', 'Nevada']**

**Outlier states for UrbanPop : []**

**b.Z-score Method**

**The Z-score method identifies outliers by calculating how many standard deviations each data point is away from the mean of the dataset using the formula Zi​=σxi​−μ​. A predefined threshold (typically 3) is then applied to the absolute Z-scores; data points with Z-scores exceeding this threshold are flagged as outliers, indicating they are significantly distant from the average value, assuming a roughly normal distribution of the data. This standardization allows for a consistent measure of unusualness across the dataset.**

****

**Outliers in Murder (|z-score| > 2):**

**Location Murder**

**9 Georgia 17.4**

**Outliers in Assault (|z-score| > 2):**

**Location Assault**

**32 North Carolina 337**

**Outliers in Rape (|z-score| > 2):**

**Location Rape**

**1 Alaska 44.5**

**4 California 40.6**

**27 Nevada 46.0**

**Percentage of outliers in each variable**

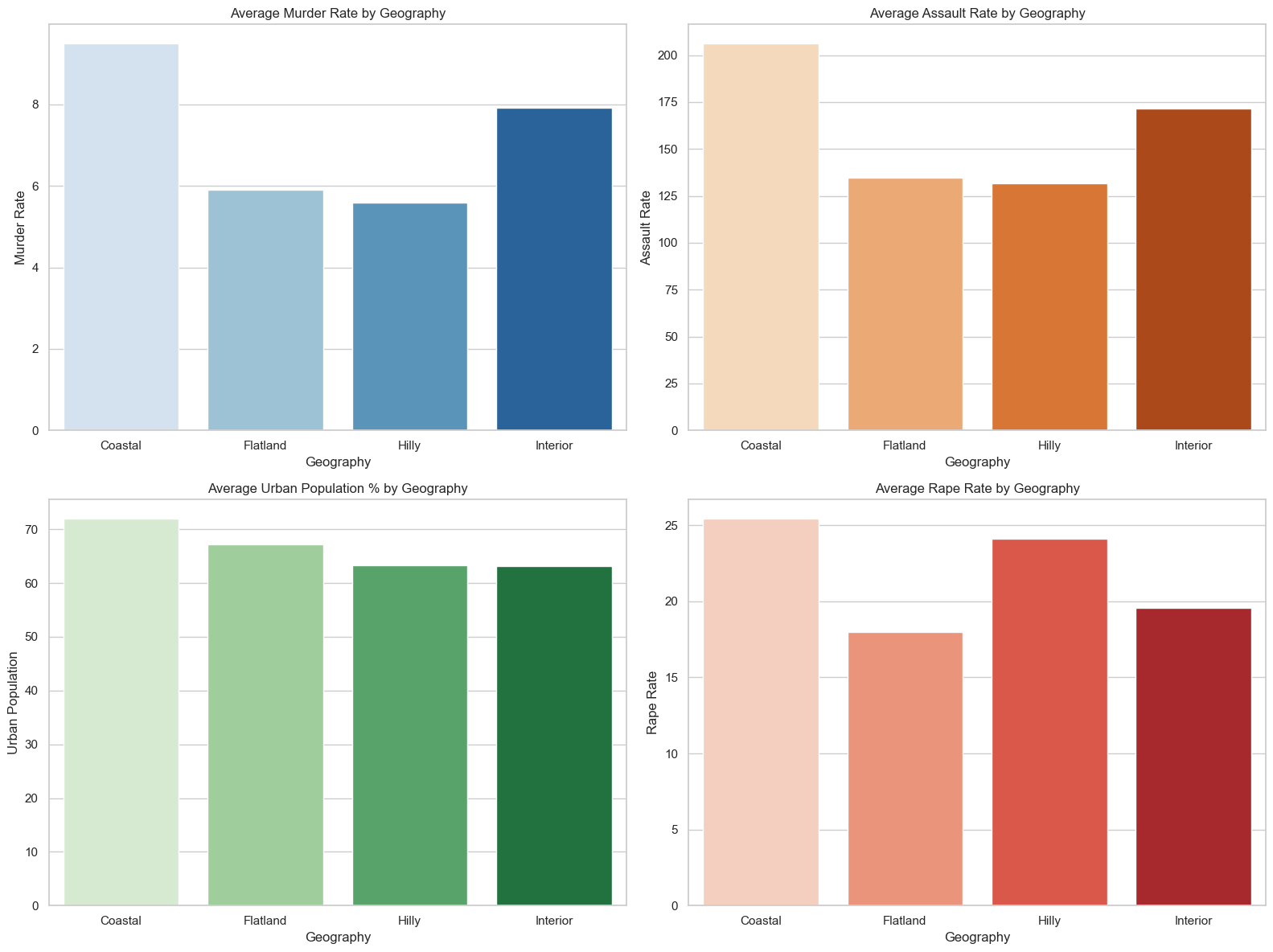
**Percentage of outliers in each variable:**

**Murder: 2.0%**

**Assault: 2.0%**

**Rape: 6.0%**

**7.Crimes based on Geographic Relief Features(Used Folium)**

****

| **Geography** | **Murder** | **Assault** | **Urban Population** | **Rape** | **Insights** |
| --- | --- | --- | --- | --- | --- |
| **Coastal 🌊** | **Moderate** | **High** | **Very High** | **High** | **Urbanization drives more Assault and Rape cases; cities exposed.** |
| **Hilly 🏔️** | **Low** | **Low** | **Low** | **Moderate** | **Hilly isolated areas have fewer crimes, but personal crimes (rape) slightly exist.** |
| **Flatlands 🏜️** | **Moderate** | **Moderate** | **Moderate** | **Moderate** | **Assaults and murders can spike due to socio-economic inequalities.** |
| **Interior 🏞️** | **High** | **Moderate** | **Low-Moderate** | **High** | **Rural isolation leads to slightly higher murder and rape cases.** |