**Data Processing:**

### Dataset Summary:

* The dataset was imported from a csv file, containing 250 values, with 5 columns and 50 records.The columns include:

| **Feature** | **Data Type** | **Description** |
| --- | --- | --- |
| City | String | The column represents all the states in the United States of America |
| Murder | Float | The murder rate within the city |
| Assault | Integer | The total number of assaults within the city |
| UrbanPop | Integer | The total number of within the population |
| Rape | Float | The murder rate within the city |

### Data Cleaning:

1. **Duplicated Values:**
   1. Identify the total number of duplicated values by chaining the ***.duplicated()*** method and the ***.sum()*** methods to return the total number of blank values within the dataset.
   2. There are no duplicate values in the dataset.
2. **Null Values:**
   1. Identified the total number of null values by applying the numpy function ***np.count\_zero()*** and passed the dataframe with the method ***.isnull()*** chained with the ***.sum()*** method into the function.
   2. There were no null values within the dataset.
3. **Outlier Values:**
   1. The outliers were visualized using the seaborn library’s function, ***sns.boxplot()***
   2. The user defined function, ***outlier\_imputation().*** This function implemented the IQR to find the total number of outliers in each numeric column. This would then be outputted to find
   3. The same function was then responsible for imputing the outlier values, by either the mean, median, or upper and lower limits that were defined within the function.
   4. The distribution of the features were then visualized using the ***sns.histplot()*** to compare the distribution of the feature before the imputation, and after the imputation

### Data Transformation:

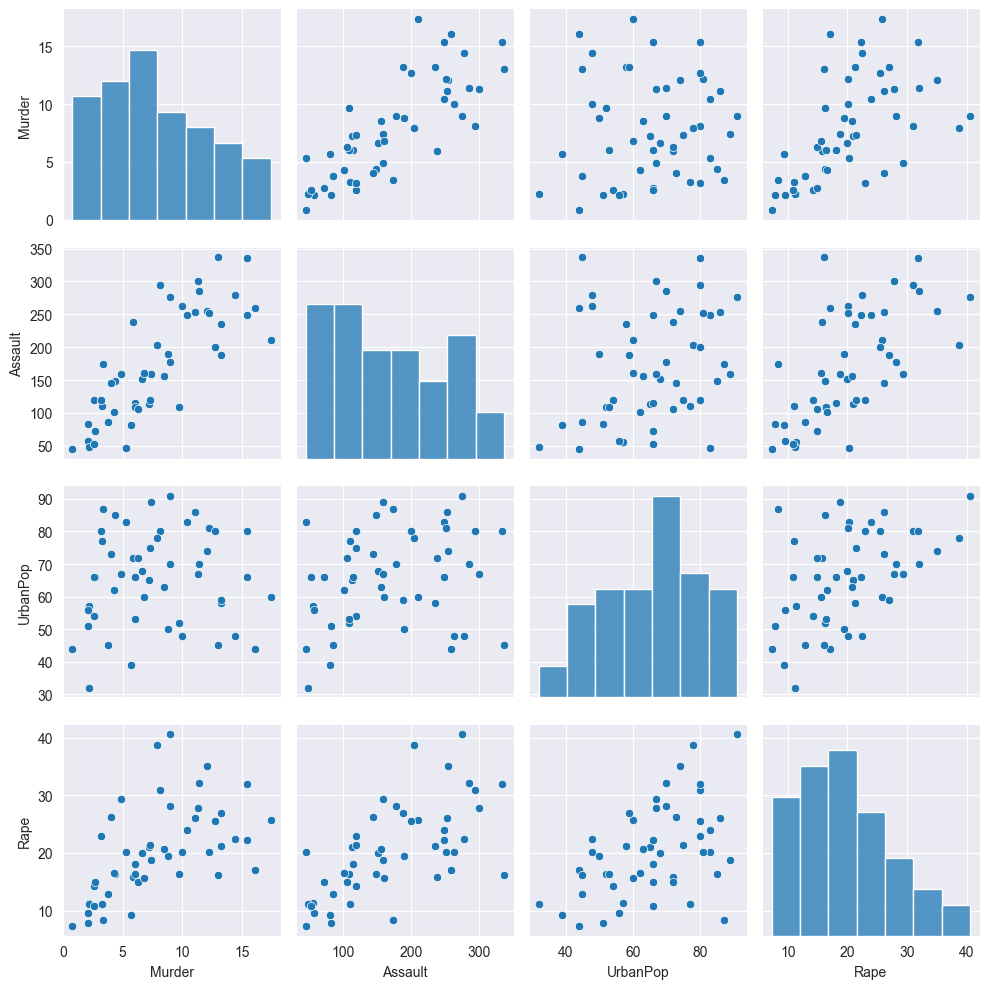
1. **Data Normalization:**
   1. Applied Min-Max normalization to scale numerical features to the range [0, 1].

### Data Engineering:

1. **Feature Engineering:**
   1. Read in geospatial data, giving information about the geographical shape of a region, in this case, the states of America
   2. The original dataset did not have a feature that we could link to the geospatial dataset.
   3. Read in the dataset scraped from wikipedia about the state, region and state code, which matches the codes of the states in the original dataset stored as df.
   4. Filter the webscraped dataset from wikipedia stored as state\_code\_df to only include the state column and the state code, and store as dictionary, in the variable state\_code\_dct.
   5. Store a copy of the original dataframe using the ***.copy()*** method. As df3.
   6. Add a new feature, Code, into df3, based on the column ‘City’. Apply ***.map()*** and pass state\_code\_dct as argument so the value from the dictionary is returned based on the value in the ‘City’ column matching the key of the dictionary

### Data Stories and Visualizations:

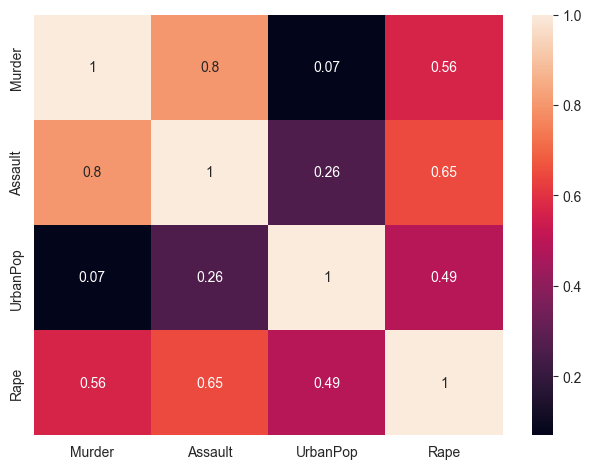
**Relationships and Distributions:**



In an article on *Geeks For Geeks (2024)*, the purpose of the pairplot is to visualize the relationship between the numeric variables and distribution of the numbers in the feature.

Some interesting information gleaned from the relationship between the murder and urban population features is that there is no correlation. A correlation matrix can be used to return a more clearer picture on the features relationships.

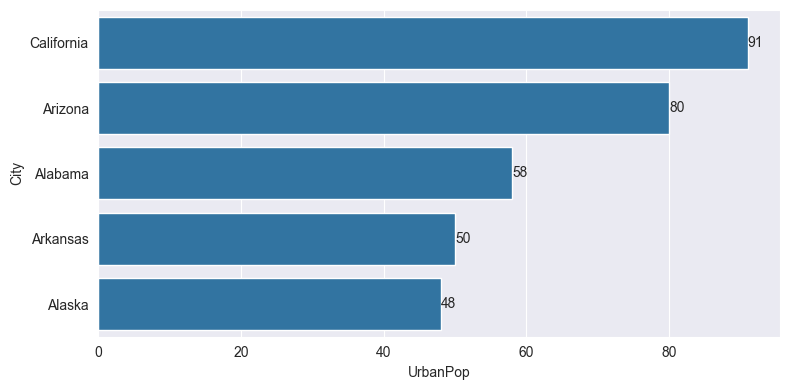
We can also see from the graph that there is a strong correlation between the murder and assault featuers for each city.

**Correlation:**

The initial assumption is that there is no correlation between urban population and the violent crimes within the each state

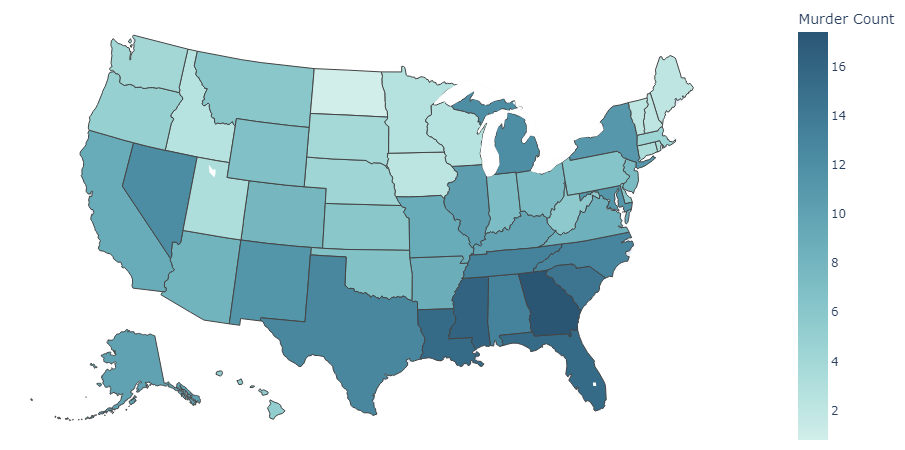
There is an assumed strength in the correlation between the murder and assault features. In his book, Data Simplification, *Jules J. Berman (2016)*, explains that a pearson correlation coefficient of 0.8 or greater represents a strong relationship between the two features.

**Top 5 states per population:**

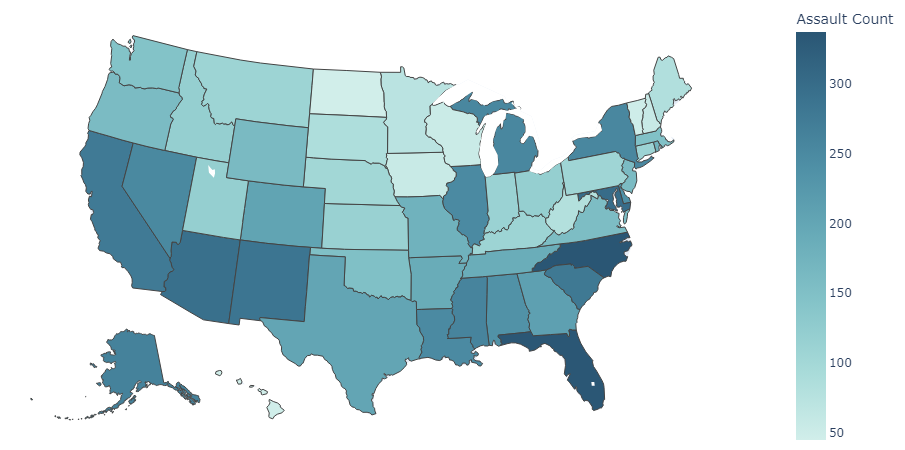
Based on the bar chart above, we can see that the top 5 states from the dataset includes: California, Arizona, Alabama, Arkansas and Alaska

**Relationships and Distributions:**

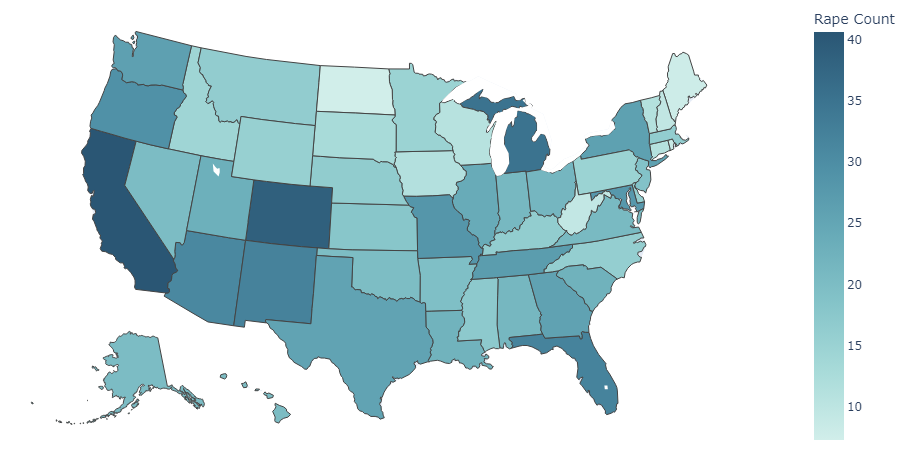
*Murder Count:*



*Assault Count:*

**

*Rape Count:*

**

Based on the choropleth maps, we can see that the highest concentration of murders occur in the south east of America where you will find states such as Louisiana, Mississippi and Georgia. Coincidently, Regionally, the highest concentration of Black people in the U.S. in 2022 is in the South; more than half (56%) live there. (Moslimani, 2024). However, seeing two correlated features does not necessarily mean that we can infer that one variable causes another.

We can also see that there is a more spread area when it comes to Assault crimes as there is a greater concentration across the south of America, with Places such as California, Arizona, and New Mexico in the west, adding to the States in the East with their larger numbers comparatively.

Interestingly, we find that the highest Rape numbers come from the western coastline.

**Hypothesis Testing:**

Based on the null hypothesis, we assume that there is no linear relationship between the murder rates and the assualt cases in the states of America.

The alternative hypothesis will therefore be that there is a linear relationship between the murder rates and the assault cases in the staes of America.

We set the significance level at 5%. Therefore, based on the pearson coefficient, we find the P value using the t\_test using the ***pearsonr()*** function from ***scipy.stats*** library. If the p-value falls below the 5% level, this would lead us to reject the null hypothesis, and accept the alternative hypothesiis, which in this case is that there is a linear relationship between the murder and assault features of the dataset.

**Recommendations:**

Further research should be conducted to find the factors that could be the cause for the increased violent crimes amongst the most prevalent states for each violent crime, and compare the factors to the states that have a lower rate of violent crimes. The results of said research could then be presented with conclusive evidence, to governments of states to implements policies that could decrease violent crimes and increase the safety and quality of life for the residents of the states.

**Bibliography:**

1. \_Geeks For Geeks\_. (2024), accessed 11 Jun 2024, Accessed at:

<https://www.student.unsw.edu.au/how-do-i-cite-electronic-sources>

2. Jules J. Berman. (2016) \_Data Simplification\_, Morgan Kaufmann is an imprint of Elsevier, Cambridge, MA, Available at:

<https://www.sciencedirect.com/topics/computer-science/pearson-correlation#:~:text=Pearson%20Correlation%20is%20a%20statistical,near%20zero%20indicates%20no%20correlation.>

3. Analytics Vidhya. (2024), \_Difference Between Skewness and Kurtosis\_, 29 July 2024, Accessed at:

<https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/>

4. Zakaria Jaadi. (2023), \_'When and Why to Standardize Your Data A simple guide on when it is necessary to standardize your data.'\_, 29 July 2024, Accessed at:

<https://builtin.com/data-science/when-and-why-standardize-your-data#:~:text=Normalization%20is%20useful%20for%20when,is%20useful%20for%20normal%20distributions.>

5. GitLab Documentation. (no date) \_GeoJSON files\_. Available at:

<https://docs.gitlab.com/ee/user/project/repository/files/geojson.html#:~:text=A%20GeoJSON%20file%20is%20a,along%20with%20their%20associated%20attributes.> (Accessed: 01 August 2024)

6. Moslimani, M. (2024) \_Facts About the U.S. Black Population\_. Available at:

<https://www.pewresearch.org/social-trends/fact-sheet/facts-about-the-us-black-population/#:~:text=Regionally%2C%20the%20South%20is%20home,the%20total%20as%20of%202022.> (Accesed: 01 August 2024)