Hyperiondev

TASK

Exploratory Data Analysis on the USA Arrests Dataset

Visit our website

Introduction

Summary of the data set

The 'UsArrests.csv' file is a text file containing information about statistics in arrests per 100,00 residents for assault, murder, and rape in each of the 50No. US States in 1973. Also given is the percentage of the population living in urban areas.

- City – Full USA State name (String)
- Murder - Murder Arrests per 100,000 residents (Float.),
- Assault - Assaults arrests per 100,000 residents (Int.),
- Rape - Rapes arrests per 100,000 residents (Float.),
- UrbanPop - Percentage population per state living in Urban Environment (%) or inferred population density for given State.

There is a total of 50No. records and 5No. unique feature classes.

I have assumed that crime rate values are representative of that State inside and outside of urbanised areas. I.e., we don't really know how the crime rates are reflected between urbanised and non-urbanised area.

DATA CLEANING

The following methods have been used in cleansing the dataset, where appropriate visualisations are included.

1) Head() – To show data rows and columns, gaining an appreciation of value data types.
2) Info() – To show a summary of the dataframe, including the number of rows and columns, the column names, and datatypes.
3) Rename() – To rename 'City' feature class to 'State'.
4) Missingno.matrix() visualisation – From observation, no values appear to be missing.

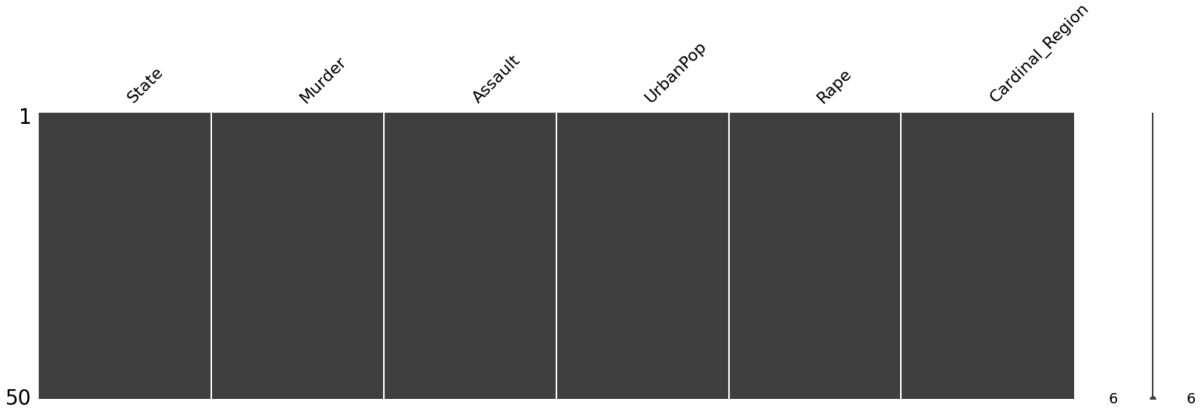

Figure 1: Missingno Matrix (null values)

5) Astype() – To allow for full assessment, price, horsepower, peak-rpm, stroke, bore and engine-size data values cast from object class to integer.
6) Describe() – Reviewing the initial data shows 'Assault' has a high variance which indicates to us that in comparison to our other variables, we should consider scaling all data to make comparative observations. The scale in magnitude is understandable given common sense comparisons in the type of crimes these variables represent i.e. Assault be the most common and murder being the least common in terms of arrests per 100,000 residents.

Data Transformation

1) Minmax scaling() – Due to differences in magnitude and distribution, continuous data columns have been scaled to compare data with different magnitudes and ranges so that one feature does not dominate the other feature. These scaled data columns are added to the main dataframe.

Data Addition

1) Given full state name, adding additional columns relating to that geographic data could be enlightening in making wider inferences about these types of crimes. For example, it may be possible make lists of set states by geographical items inter alia, region, political vote, or temperature.
2) A 'state_codes' dictionary was formed relating key and value pairs by full state name and state code.
3) A dictionary of geographical regions created called 'cardinal_regions' that includes lists of 'north_east', 'north_central', 'south' and 'west' state codes.
4) A function was created to assign a region name to each record in the dataset based on its 'State' value. The function takes each records 'state' value, searches for its value within the 'state_codes' dictionary, then proceeds to loop through 'cardinal_regions' dictionary searching for the state code within each region list.

```
# Geographical Regions.
north_east = ["PA","NJ","NY","CT","RI","MA","NH","VT","ME","HI",]
north_central = ["ND","MN","WI","MI","SD","NE","KS","IA","MO","IL","IN","OH",]
south = ["OK","TX","AR","LA","KY","TN","MS","AL","WV","VA","NC","SC","GA","FL","DE","MD"]
west = ["WA","MT","OR","ID","WY","CA","NV","UT","AZ","CO","NM","AK"]

# Dictionary to store the region as keys and lists of state as values.
cardinal_regions = {"North East": north_east,
                    "North Central": north_central,
                    "South": south,
                    "West": west}
```

Figure 2: 'cardinal_regions' Dictionary List

DATA STORIES AND VISUALISATIONS

Exploratory data analysis has been used to explore various aspects of the USArrests dataset, such as:

- Which states have the highest arrests per crime variable?
- Is there correlation between urbanisation and the number of arrests?
- What states have the highest arrests per crime above the national average?
- What is the distribution of crime rates among the states in the dataset?
- How does the crime rate vary across different regions in the USA?

Different visualisations have been used to easier understand and interpret what the data is representing.

Patterns and Trends

1) **Describe() & Data Distribution with Histograms** – To show basic statistics of the numerical columns in the dataframe and review frequency distribution. From this we can review which numerical data is included and look at the shape of the data included in the dataset.

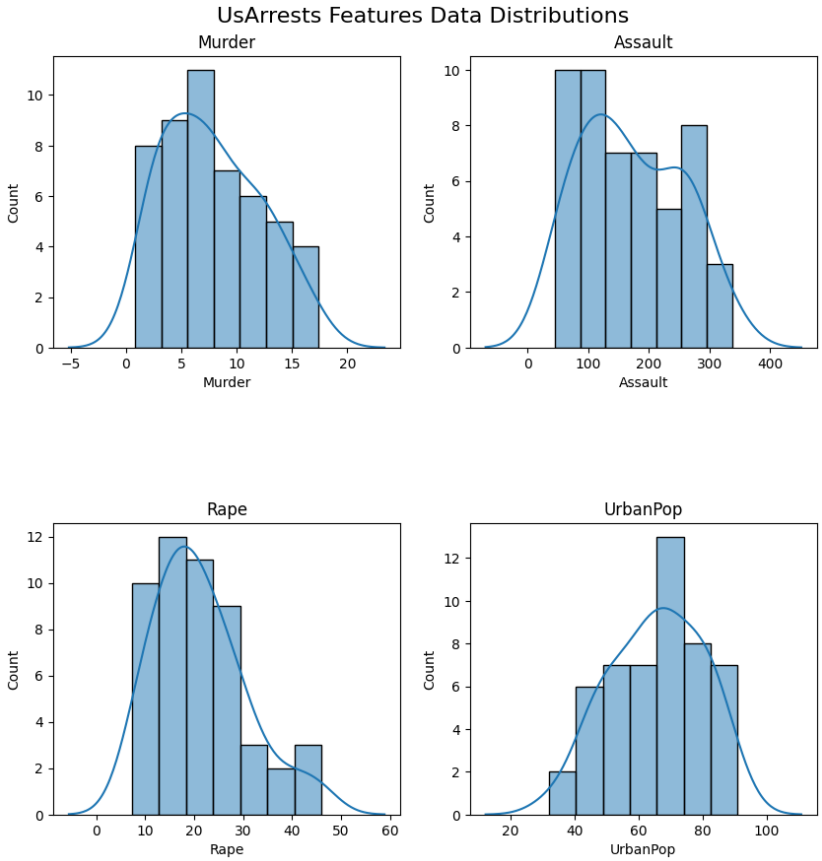

Figure 3: 'cardinal_regions' Dictionary List

- The positive skew of Murder and Rape suggests that these two forms of arrests have lower count value densities, with fewer states have higher values of murders and rapes.
- Assault is slightly positively skewed tending to being uniform, suggesting the count value densities are similarly spread there is no concentration of assaults at any specific value.

1) Describe() & Data Distribution with Histograms (Continued 1)

- UrbanPop appears slightly negative skew tending to being uniform, suggesting the count value densities are not necessarily concentrating. However, we do have a concentrates range of values at around 70%

2) Scatterplot Matrices – Pairplots were used to quickly review and visualise the relationships between UrbanPop and the arrest types in the dataset. The matrices were split by categorical 'Cardinal_Regions'.

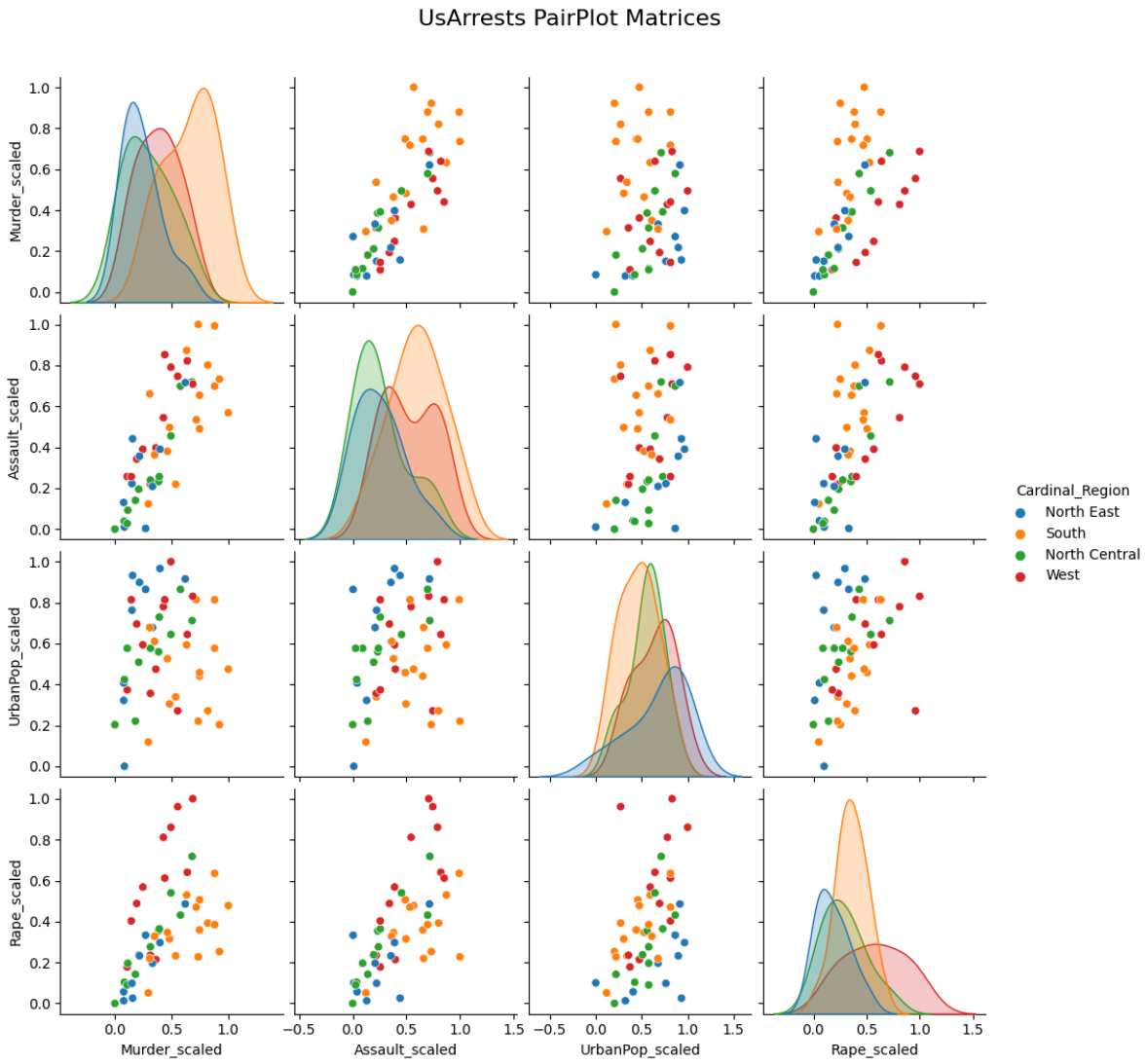

Figure 4: Scatterplot Matrix Scaled Features

2) **Scatterplot Matrices (Continued 1)**

- The data suggests a correlation between Murder with Rape and Assault. This means that an as this variable increases it is likely that the other variable will also be increasing i.e., a state with a higher murder rate may also see high rape or assault arrests.
- No correlation is observable from review of murder and assault when compared to UrbanPop. However, some weak positive correlation is visible between Rape and UrbanPop. At this stage it is not yet possible to determine this as a principal relationship, but an initial hypothesis may relate urban environments influencing the circumstances in which rapes happen. Across all scenarios, geographical comparisons may be influencing factors in such that specific states and cities have better policing. Further investigation on these points and additional data is required.
- Initial comparisons between regions highlights the Southern States generally have the highest density any crime rate (not necessarily the largest maximum crime rate value recorded) with southern states generally being tighter in distribution – This is more observable in the Rape distribution. Other regions are relatively similar in distribution except for Rape in the West region, highlighting that there is some variance in individual rape arrest rates in Western states. Identification of these states may explain further high rape arrest rates.

3) **Correlation Heatmap** – A heatmap was used to quickly review correlations between continuous feature classes.

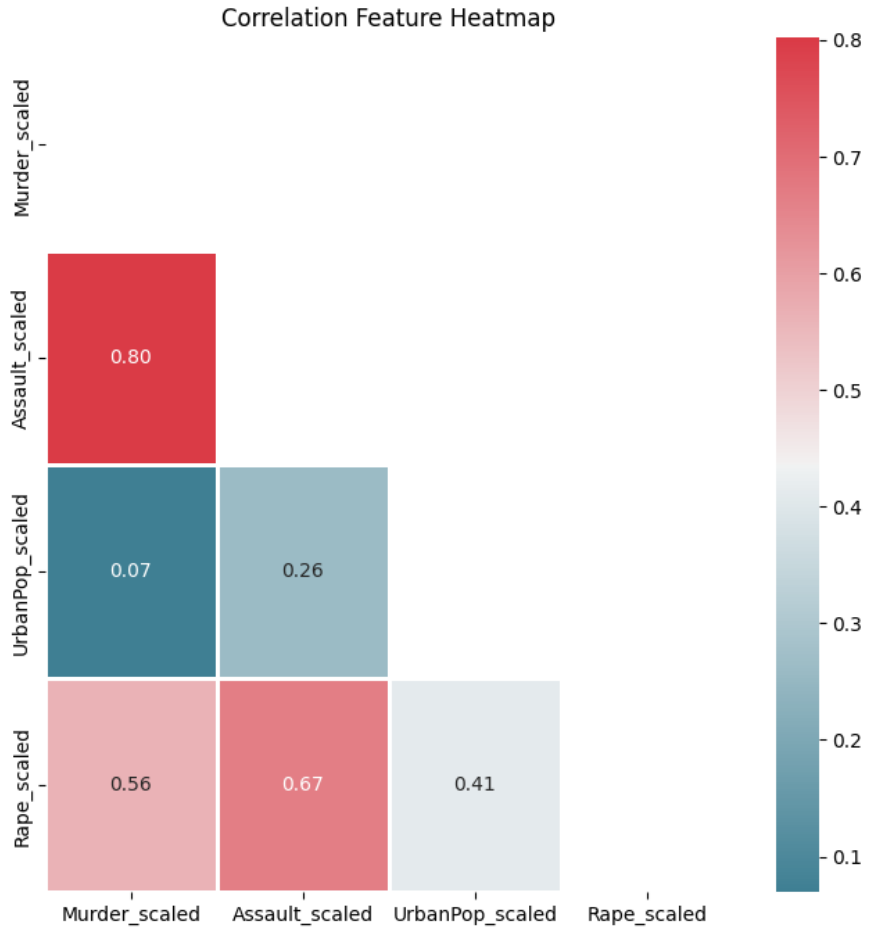

Figure 5: Correlation Heatmap of Scaled Features

3) Correlation Heatmap (Continued 1)

- Concerning correlations, we can see that murder and assault vary with each other and correlate somewhat strongly. The amount of correlation between Murder and Assault is strong here in comparison to 'Murder and Rape' and 'Assault and Rape'. For developing further visualisations, I will consider using either murder or assault to explore underlying patterns and potential categories.
- We find the weakest correlation in the dataset being ''Murder and UrbanPop', where we also find 'Rape and UrbanPop' being the strongest of the UrbanPop comparisons. As discussed above for 'Rape and UrbanPop' this pattern may be influenced by urban environments. Murder rates in comparison are much more uncorrelated and independent of UrbanPop therefore highlighting that urbanisation does not necessarily lead to an increase in murder crime rates. Other factors are probably more likely to affect murder crime rate greater such induvial state laws, and other crime rates not included in this dataset.

4) 3D Projection – A 3D Scatterplot is used to show patterns in 'UrbanPop, Assault and Rapes'. Lineplots were used to individually review X vs. Y and X vs. Z axes.

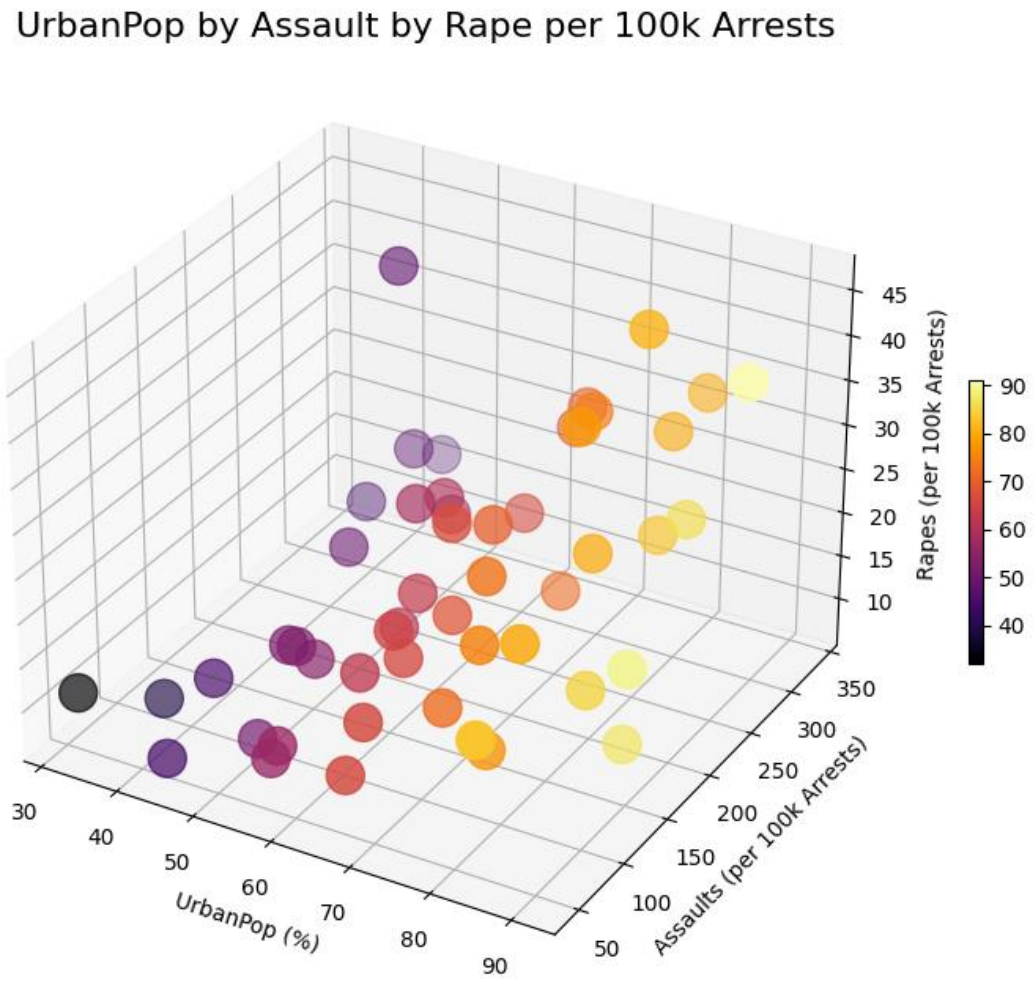

Figure 6: 3D Projected Scatterplot of Scaled Features

4) 3D Projection (Continued 1)

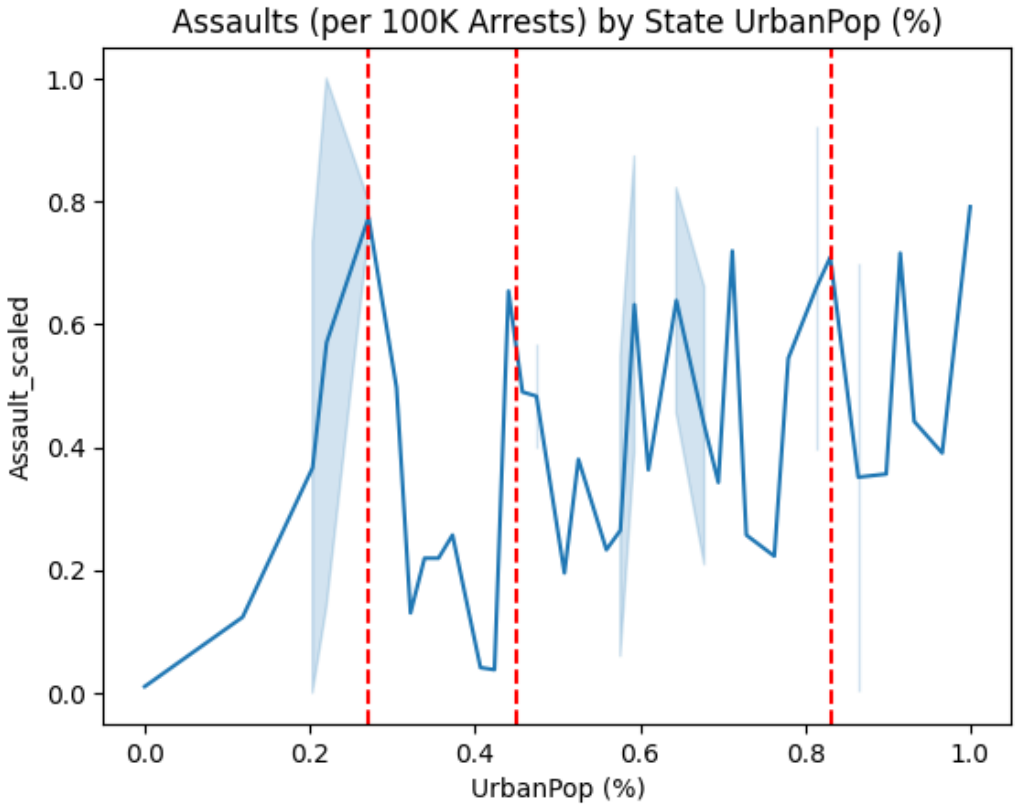

Figure 7: Lineplot of UrbanPop vs. Assault Scaled Features

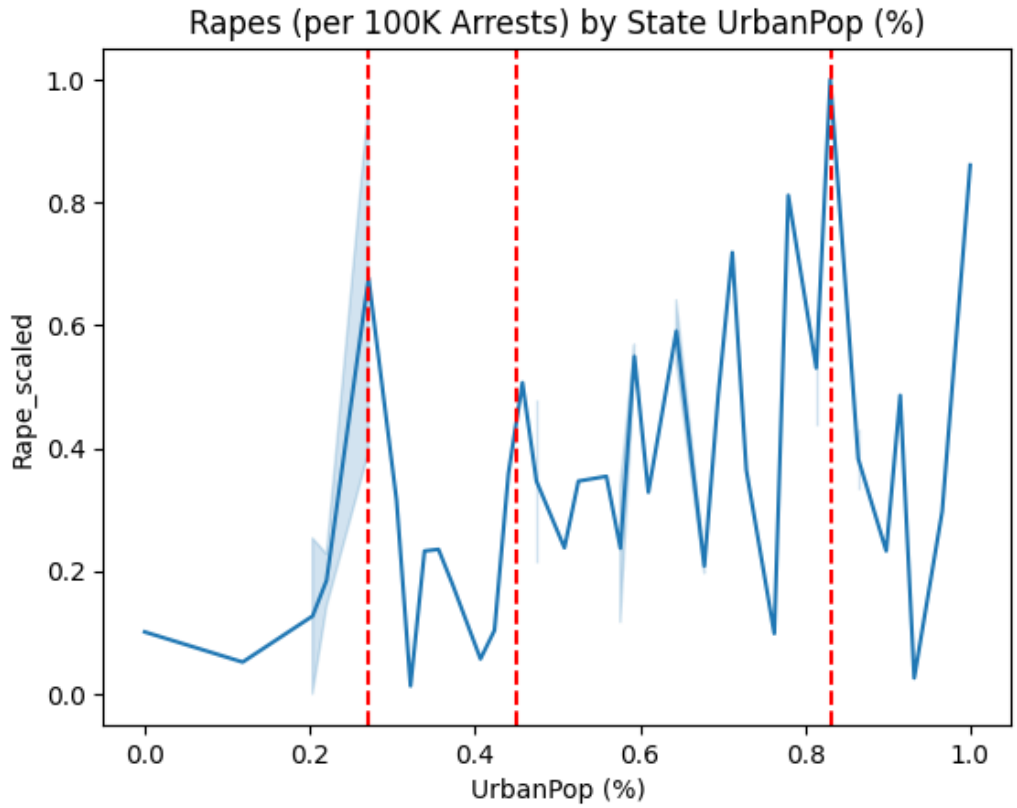

Figure 8: Lineplot of UrbanPop vs. Rape Scaled Features

4) 3D Projection (Continued 2)

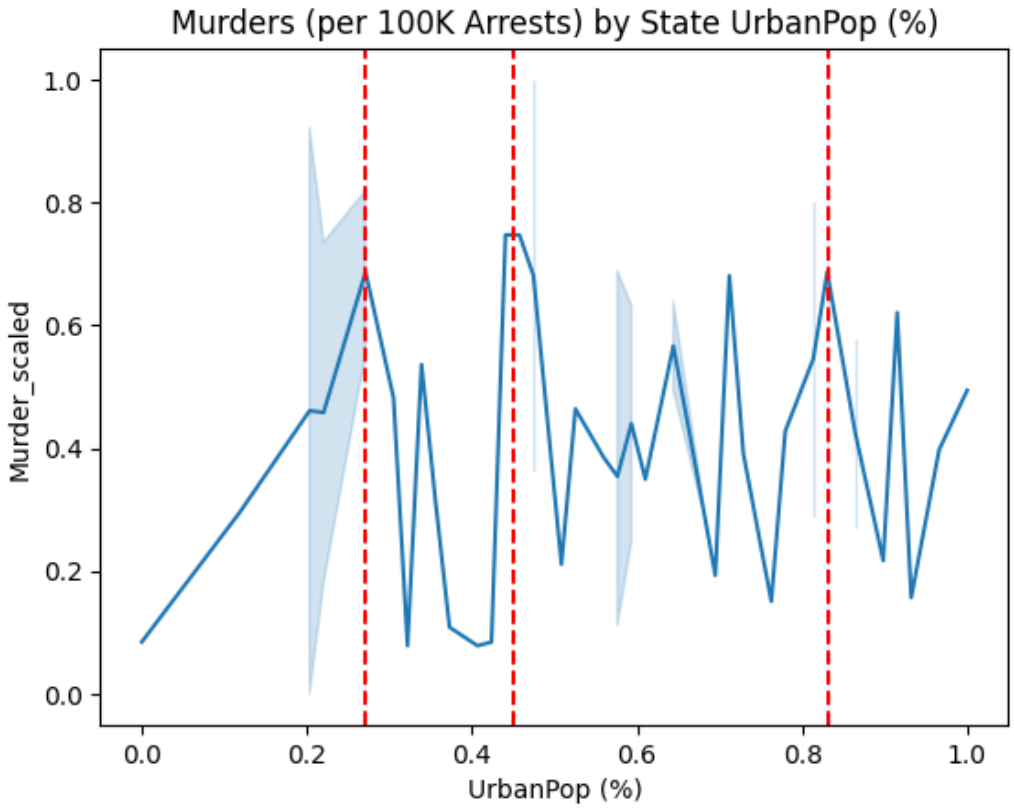

Figure 9: Lineplot of UrbanPop vs. Murder Scaled Features (Separate)

- Between Figure 6-8, a trend is discernible highlighting a general rise in assault and rape with an increase in UrbanPop.
- Exceptional cases have been traced through individual lineplots and labels (shown below) highlighting potential outliers to the observable pattern shown in the 3D Scatterplot.

5) Stack plot – A stacked area plot of scaled crime rates was created by State name in UrbanPop size order (smallest to largest). This graph is used in combination with those above.

- From the 3D Scatter plot above we began to see a trend for overall crime rates increased by UrbanPop however we did note exceptions to this pattern. Investigation into those with high crime rates by low and high UrbanPop may lead to further analysis that can be investigated.
- Looking at overall crime rates per state by UrbanPop size order all better identification of such states.

5) Stack plot (Continued 1)

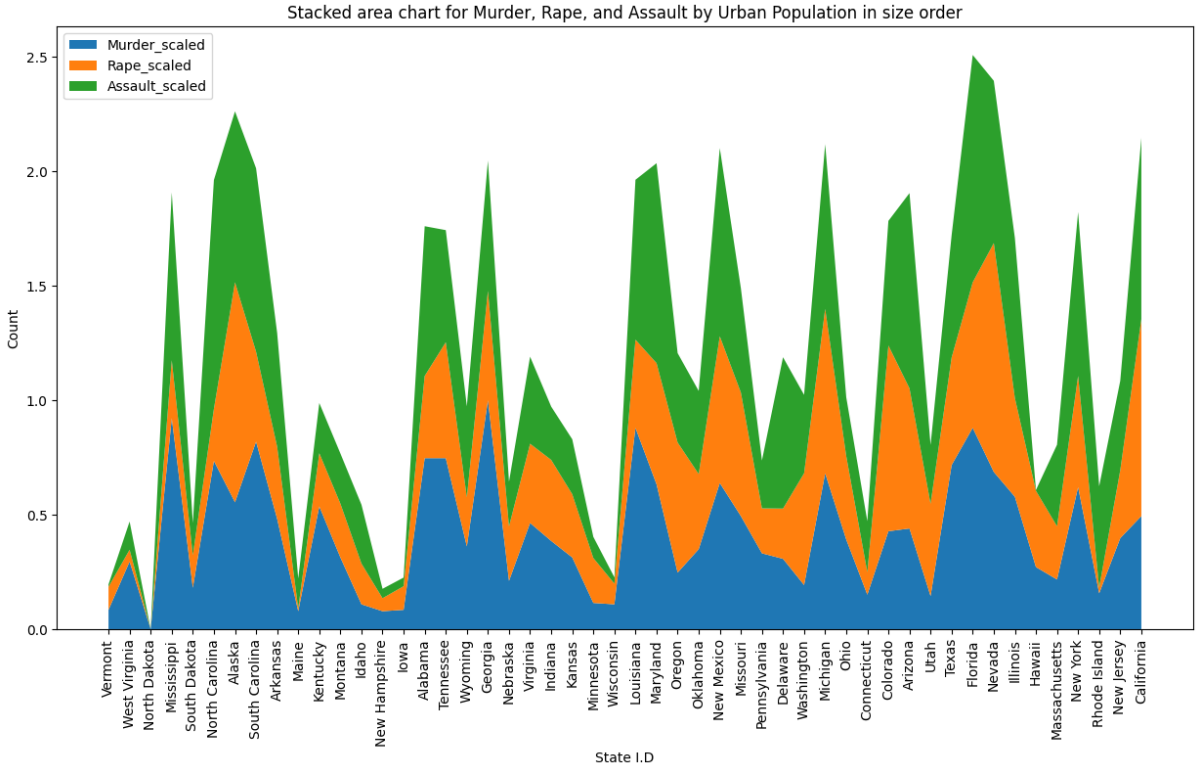

Figure 10: Stackplot of State ordered by UrbanPop vs. Scaled Features.

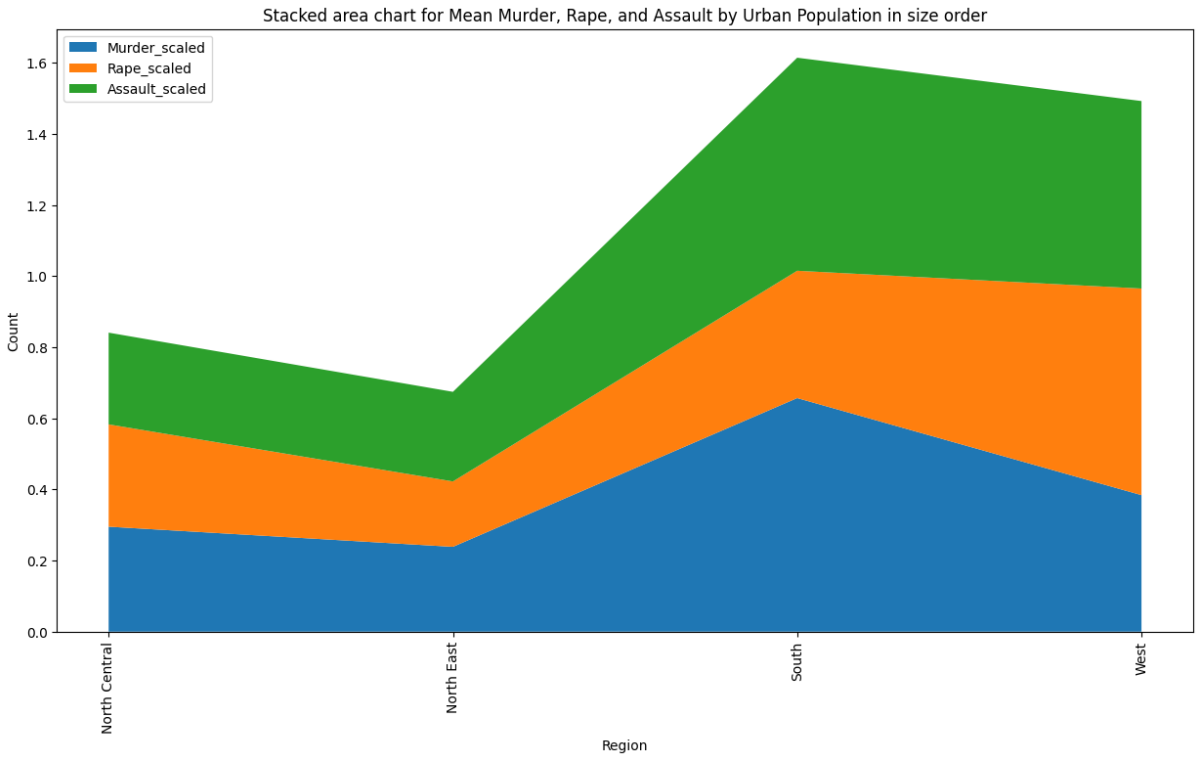

Figure 11: Stackplot of US Region vs. Mean of Scaled Features.

5) Stack plot (Continued 2)

- States with Low UrbanPop and low crime rates:
    - North Dakota (3rd lowest UrbanPop),
    - Maine (10th lowest UrbanPop).
- States with high UrbanPop and low crime rates:
    - Connecticut (37th lowest UrbanPop),
    - Hawaii (45th lowest UrbanPop),
    - Rhode Island (48th lowest UrbanPop),
- States with Low UrbanPop and high crime rates (vertical lines Graphs above)
    - Alaska (Murder, Rape and Assault, 7th lowest UrbanPop),
    - South Carolina (Murder, Rape and Assault, 8th lowest UrbanPop),
    - Alabama (Murder, 16th lowest UrbanPop),
    - Tennessee (Murder,17th lowest UrbanPop).
- States with high UrbanPop and high crime rates:
    - Florida (42nd lowest UrbanPop),
    - Nevada (43rd lowest UrbanPop),
    - New York (47th lowest UrbanPop).
    - California (50th lowest UrbanPop),
- Of interest is 'Low UrbanPop vs High UrbanPop and high crime rate,' reasons for these types of visual representations may be influenced by factors such as: poverty, lower employment opportunities, abuse and addiction and weaker law enforcement.
    - For Alaska it may also be relevant to the time such arrests/ crimes took place to see if long./short daylight hours is relevant.
    - Further data is required to review these and other factors such as historical poverty tables per state inter alia, median income, employment rates, abuse, and addition statistics.
- Of interest is 'High UrbanPop and low crime rate', possible reasons for these types of visual representation may be influenced by factors inter alia stronger law enforcement, respect for law enforcement, wealth and income.
    - For Hawaii, is known for being friendly with a strong sense of community so it is possible people value those relationships, particularly with law enforcement individuals.
    - Further data is required to review factors that may influence how a community respects each other including local law enforcement inter alia; police funding, education levels, community involvement metrics for law enforcement or volunteering work. Finding metric relating to how strong a community is, is likely to be more difficult as its more qualitative than quantitative.
- By defined Cardinal_Regions we begin to see some geographical trends such that:
    - For all included crime rates in this dataset, the South Region has the higher mean number, the West region comes 2nd, North Central 3rd and North East 4th (least).
    - Concerning individual crime rates, Rapes are more prevalent in the West Region, otherwise the South Region leads on Murder and Rape.

6) **Violin Plots** – To get a better sense of the distribution and density of crime rates and urban populations densities by region violins plots are presents.

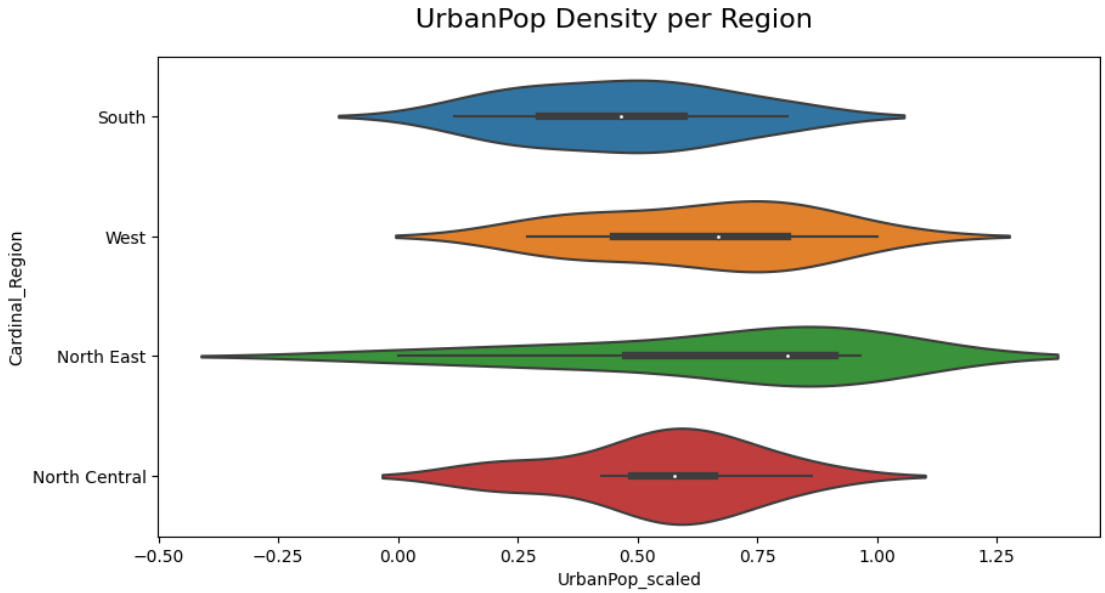

Figure 12: Violinplot UrbanPop Distributions by Region

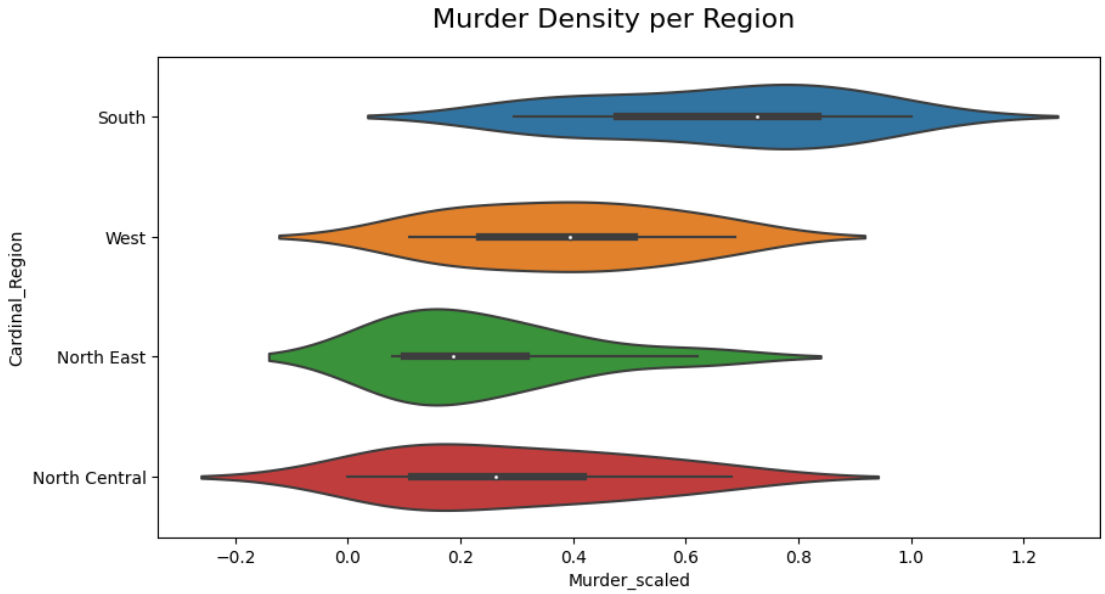

Figure 13: Violinplot Murder Distributions by Region

6) Violin Plots (Continued 1)

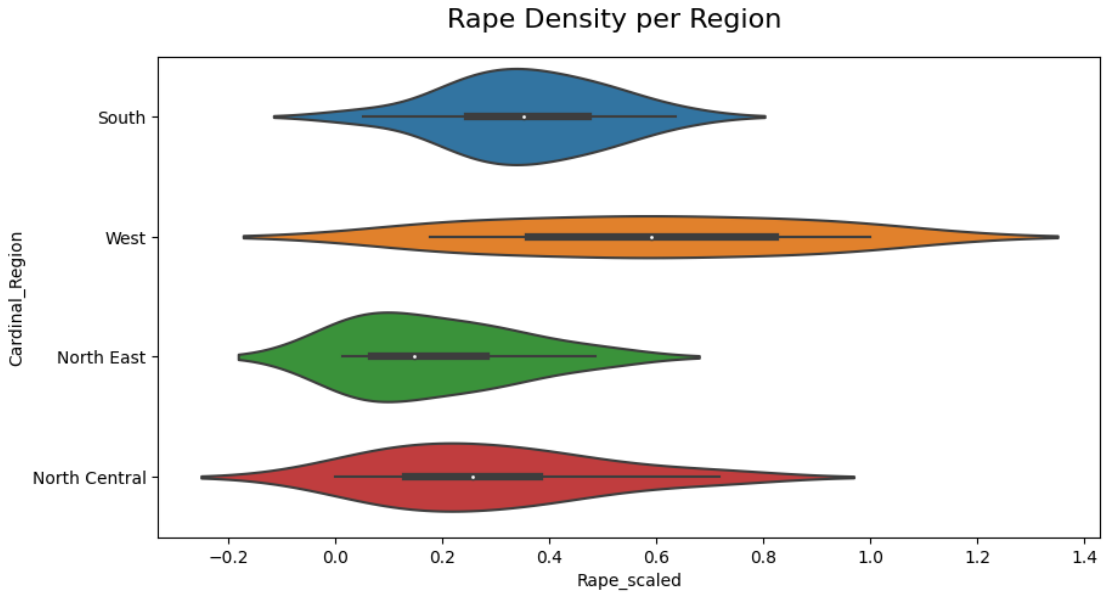

Figure 14: Violinplot Rape Distributions by Region

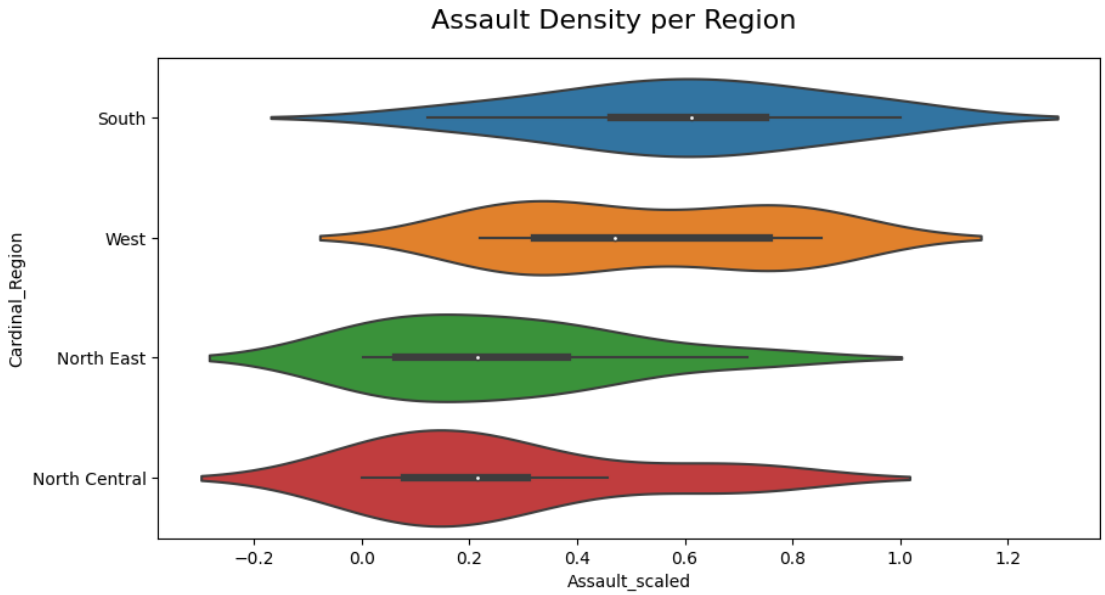

Figure 15: Violinplot Assault Distributions by Region

- Northeast Region more densely urbanised and populated however it also has the greatest variance. Regions covering the main east and west coast are most urbanised with the north central and south regions being the least.
- Concerning Murder the Northeast has the highest density of murders, with other regions being more spread out.
- Concerning Rapes, the South has the highest density of rapes, with other regions being more spread out, particularly the West.
- Concerning Assaults, densities are similar (albeit for different values of assault). Of interest is the density about the West region, being fairly uniform and not much different between high and low density areas.

7) **Bar plots** – To get a better sense of maximum value individual Murder, Rape and Assault scaled crime rates, bar plots are presented.

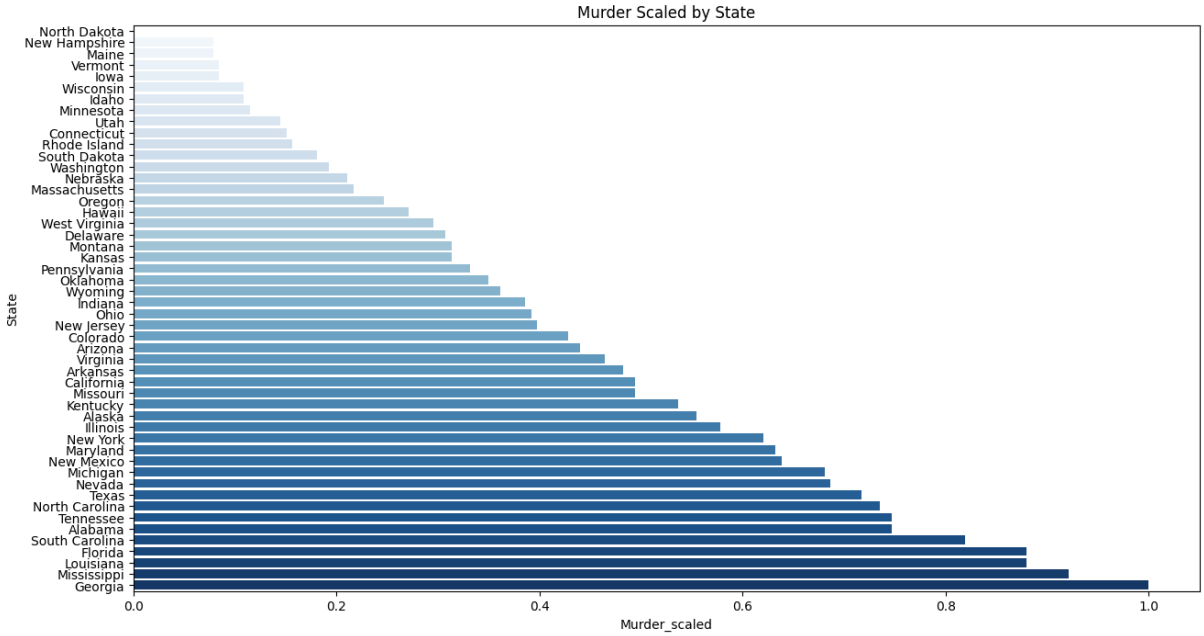

Figure 16: Barplot of scaled Murder rates

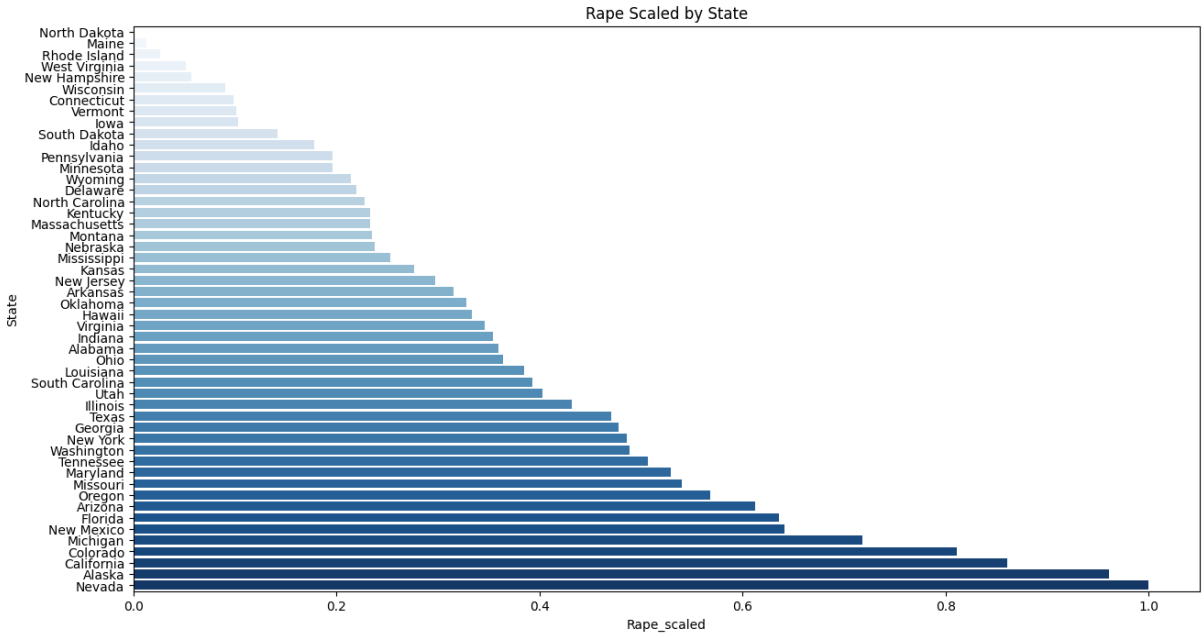

Figure 17: Barplot of scaled Rape rates

7) Bar plots (Continued 1)

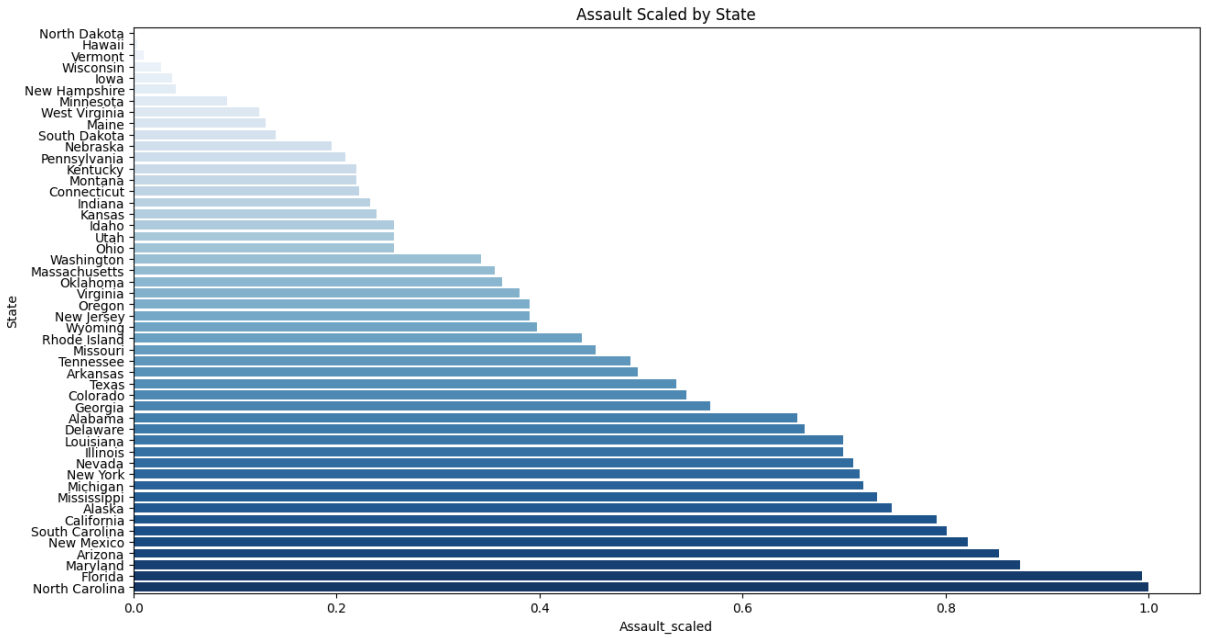

Figure 18: Barplot of scaled Assault rates

- Investigating the leading State for Rape Arrests, this state was Nevada followed by Alaska.
    - As described above for Alaska, several reasons as to why rape is more common in Alaska than elsewhere may include a higher male to female sex ratio, remoteness, attitude, or o weak police enforcement. Further investigation is required.
    - Concerning Nevada again further investigation is required. Initial factors may relate to prevalence of gambling and entertainment industry i.e., Las Vegas. Numbers for such business types could be included in future analysis to see if these types of businesses influenced types of crimes.
- Investigating the leading State for Assault Arrests, this state was North Carolina followed by Florida.
    - Finding specific reasons for higher assaults in North Carolina and Florida is difficult. More specific information is required about the condition in these two states in 1973 than is available in this dataset. Possible factors may include drug related features such as trafficking about coastal areas via the Gulf of Mexico.
- Investigating the leading States for Murder Arrests, this state was Georgia following by Mississippi.
    - Concerning Georgia, more information is required about the conditions in this state than is available in this dataset.

Main Findings from UsArrests Dataset

The USArrests dataset was analysed using exploratory data analysis techniques to better understand the relationship between crime rates and urban populations in different states in the USA. Descriptive statistics, histograms, scatterplots, correlation heatmaps, 3D projections, and stacked plots were used in the analysis.

The results showed that states in the Southern region had a higher density of crime rates, although this did not necessarily mean they had the highest maximum crime rates. The strongest correlation was between murder and rape, while the weakest correlation was between murder and urban population. An increase in assault and rape was observed with an increase in urban population, but exceptions were noted for some individual states.

Stacked area plots of crime rates by state name ordered by urban population size, as well as crime rate mean values by region, were created to identify states with high or low crime rates. Although there is a general trend of higher crime rates associated with higher urban populations, this relationship is not always straightforward. Other factors, such as the availability of drugs, economic opportunities for crime (poverty concentrations, employment), effective policing, and the socioeconomic status of the community, are hypothesised to play a role in determining crime rates.

Further analysis of the states with high and low urban populations and crime rates would require additional data and a more in-depth analysis to understand the reasons behind the trends. The current dataset is limited by only including population densities by state and state name.

THIS REPORT WAS WRITTEN BY : KARL GIBSON