Spring Quarter, 2021

Sameer Patel

University of Denver

Analyzing Crime Data via K-Means Clustering

Intent: Can clustering methodologies, and associated visualizations, be used to determine relationships between features of an unlabeled dataset?

# Introduction

The intent of this analysis is to use clustering methodologies and data science visualization tools to determine whether relationships exist in an unlabeled dataset containing crime statistics for the 50 states in the U.S (1973).

This exploration will consist of the following:

* Loading the data
* Exploratory data analysis (EDA)
* Data preprocessing
* Model building
  + Hierarchical Clustering
  + K-Means Clustering
* Model evaluation / visualizations
  + Pairplots
  + Geographic chloropleth
* Conclusion / future analyses

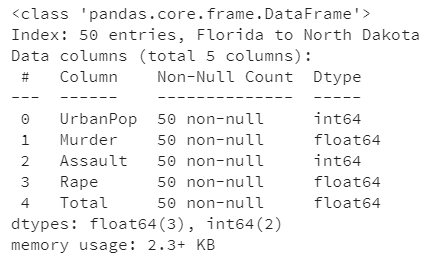
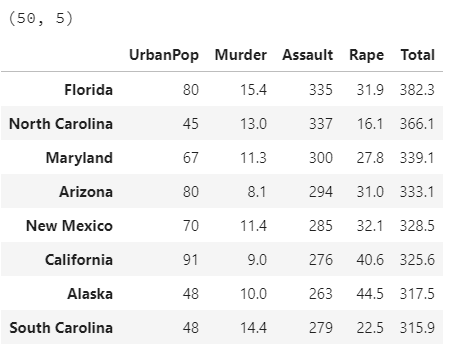
The conclusion will assess the efficacy of determining relationships amongst the data using clustering and derive further routes of exploration based on any questions that might remain once the analysis is completed.

# Dataset Description

Data is sourced from <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/USArrests.html>.

This data set contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percentage of the population living in urban areas.

Below is a snippet of the full dataset, along with information about datatypes and non-null count for each of the features.



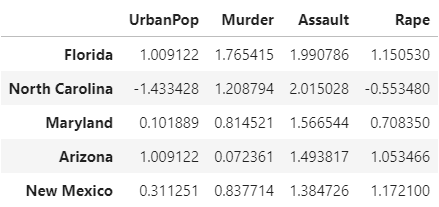
# Data Preparation

Not much preprocessing is required for this particular dataset, as all the features are numeric and there are no missing values.

Most applications of clustering methods in data science and machine learning require the features to first be scaled prior to being used as inputs to the model. This intuitively makes sense, as the scales of each of the features could vary from one-another and could therefore have imbalanced influences on the resulting clusters.

For this exploration, we will make use of the preprocessing library in sklearn, specifically StandardScaler(), in order to scale our input columns prior to clustering.

Below is a snippet of the scaled dataset.

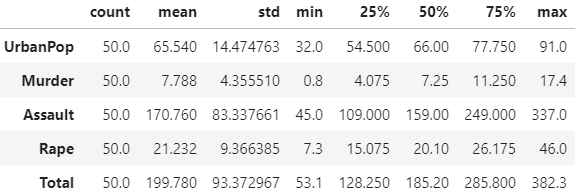


# Data Analysis (EDA)

First, we will create an overlaid bar plot of the total crime and percentage of urban population by state to get a feel for the dataset. We will also create a stacked bar plot to show proportionality of the crime types in the dataset. In the interest of space, these visualizations can be found in the appendix (Figures 1 and 2).

Based on a cursory look at the stacked horizontal bar plot of total crime vs. percentage of urban population, no immediate trends stick out vis a vis a relationship between the two variables. However, a slight geographic trend can be observed, as the majority of the top 20 most crime-ridden states are from the American South. The proportionality bar chart only tells us that crime-heavy states tend to have more of each crime type contained within the dataset.

The descriptive statistics presented below reveals to us that the variances of the input columns vary significantly from one-another.



A heatmap of feature correlation (Figure 3) and a pairplot of the features (Figure 4) indicates that all the features have some positive correlation with all the other features, with the strongest-correlated features being Murder and Assault.

Perhaps most surprisingly, however, is the low correlation (Pearson's coefficient close to 0) between UrbanPop and Murder. This revelation is surprising because of the fact that many people assume murders are more likely to occur within or near urban centers, whereas the data so far seems to indicate that many people in the United States enjoy killing each other regardless of where they live.

***Note****: for clustering analyses, feature collinearity is not an issue, as clustering methodologies do not rely on linear assumptions, but rather user-defined distance metrics.*

From the KDE plots on the main diagonal we can deduce that all the feature variables are approximately normally distributed. Though normality is not a necessary assumption of K-Means Clustering, the model is sensitive to differences in variance of the input features, confirming the need for scaling prior to model fitting.

# Data Analysis (Model Building)

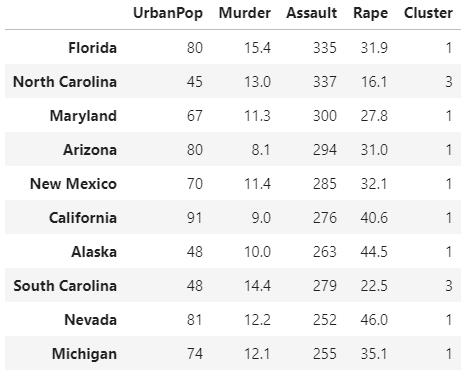
We are now ready to build our K-Means clustering model on the processed crime data. In order to do this, we will first generate a scree plot (Figure 5) of the errors (inertia) based on cluster numbers in range(1,16) in order to determine the ideal input number of clusters to the model.

The scree plot (AKA elbow plot) of inertia vs. the number of clusters used as an input to the KMeans() constructor indicates that the ideal number of clusters for our dataset would be 4. This is because the plot shows an "elbow" at x=4, indicating that the reduction of inertia that can be achieved by increasing the number of clusters beyond 4 does not outweigh the cost of adding to the complexity of the model.

To confirm the resulting cluster number from the scree plot, we will construct a dendrogram (Figure 6) through the use of hierarchical agglomerative clustering to visually check the ideal cluster number.

We can see from the dendrogram that a distance of ~75 yields 4 clusters (determined by slicing the tree with a horizontal line at distance=75 and counting the intersections). No other splits are made until a distance of ~125, confirming the conclusion reached by the scree plot.

Below is a snippet of the dataframe containing cluster predictions for each record.



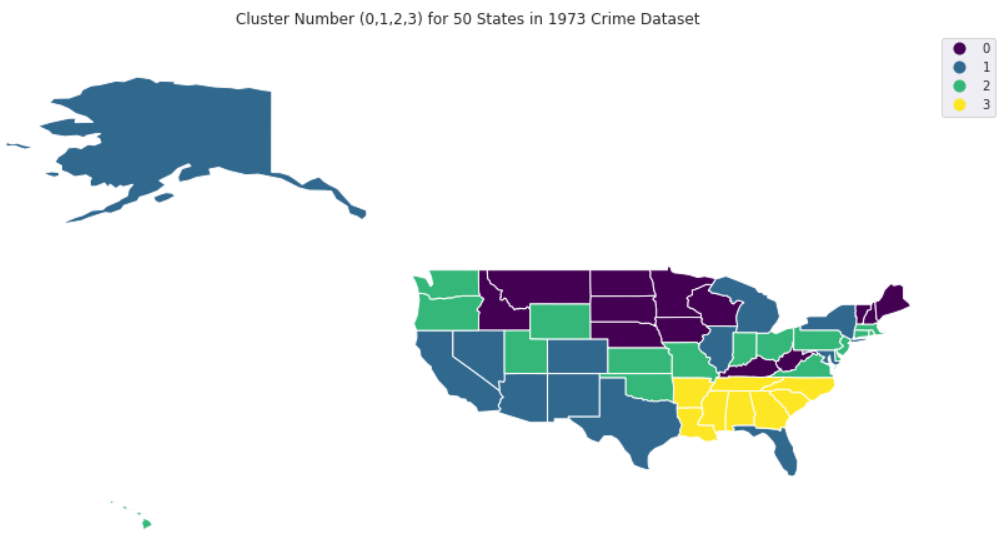
# Model Results

Because K-Means was fit to 4 feature variables, visualization of clusters/cluster centers is not doable. However, we can reproduce a pairplot with "Cluster" as the hue to visualize groupings based on pairwise features (Figure 7).

Based on the densities along the main diagonal, Murder and Assault appear to be the most influential features in determining cluster centers. Even though all of the clusters show some degree of overlap, these two features have the most distinct distributions between clusters.

A decent degree of spread and overlap exists in the pairwise scatterplots between each of the clusters for all of the variables, most notably for Rape. Again, however, the above visualization is compressing a 4-dimensional clustering model into 2 dimensions, so steadfast conclusions about the strength and validity of the model based solely on the pairwise scatterplots is not feasible.

The next step of our analysis will be examining whether a geographic relationship exists amongst the clusters. The geographic data necessary to define state geometries is sourced from COMP 4433 Data Visualization, Assignment 5 (stored as a pickle file in the directory).



# Discussion, Conclusion, and Future Analysis

Surprisingly, a very clear geographic relationship exists amongst the clusters.

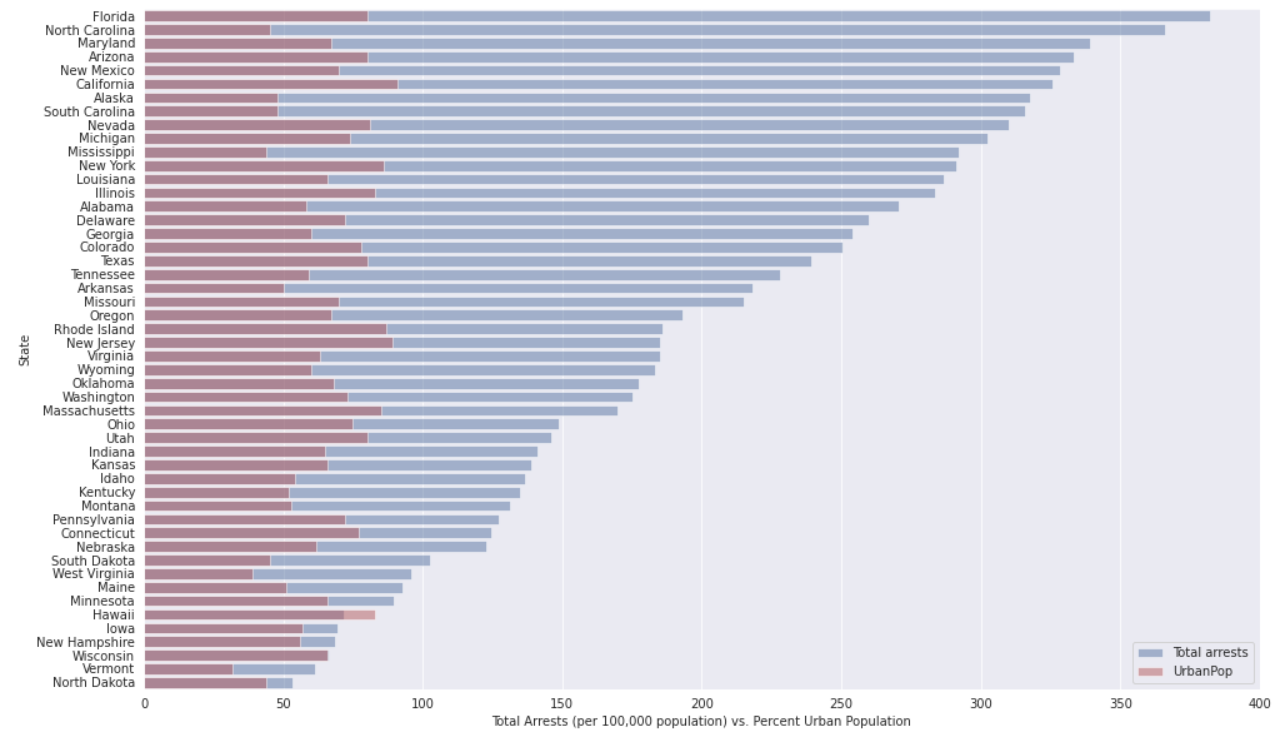
Cluster 3, which tended to have higher crime numbers across the board, is neatly localized in the Deep South. Cluster 1 also had relatively high crime numbers, which could explain the grouping of high-population states within it (e.g. Texas, Florida, California, New York, Illinois). More rural states were grouped into Clusters 0 and 2, with Cluster 0 tending to be further North than Cluster 2.

For future analyses, supplementary datasets could be added to the model, including variables such as total population in 1973, and metrics such as happiness index or state-wide GDP.

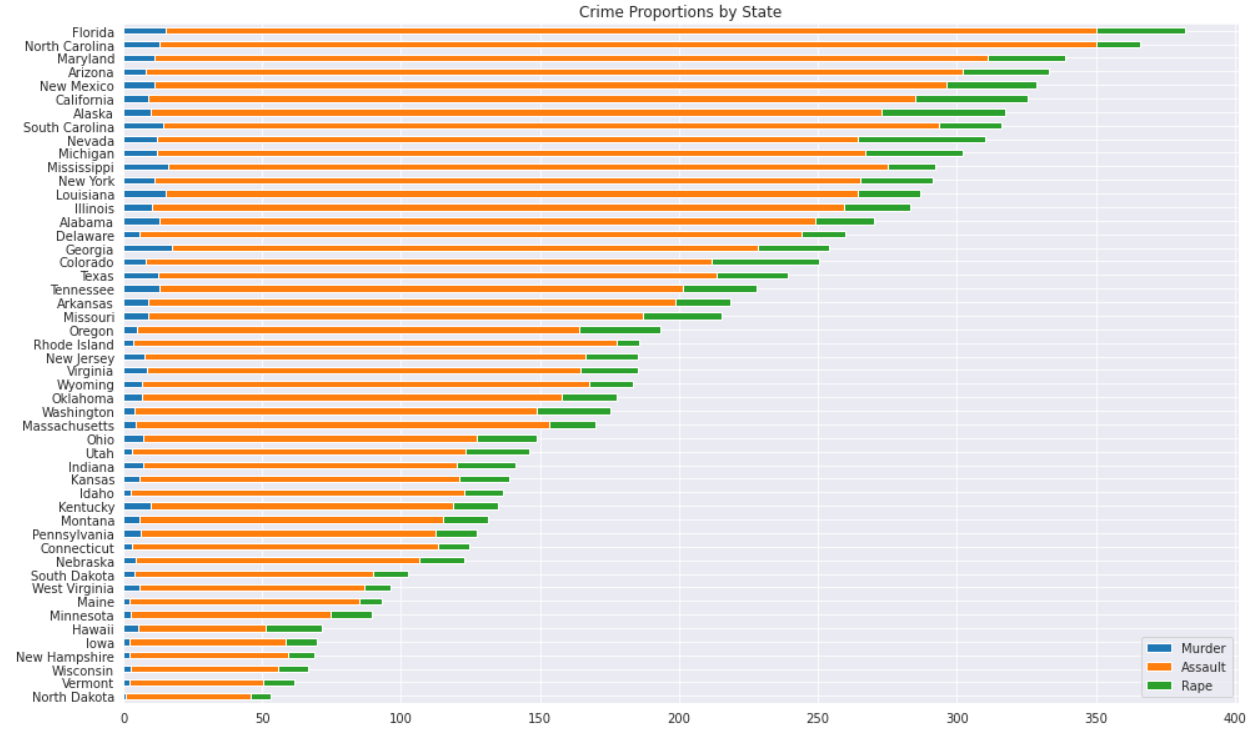
With the addition of other variables, perhaps PCA could be employed in the preprocessing section in order to reduce the number of input features while preserving the variance, and possibly leading to better cluster visualization depending on the number of returned components.

Appendix

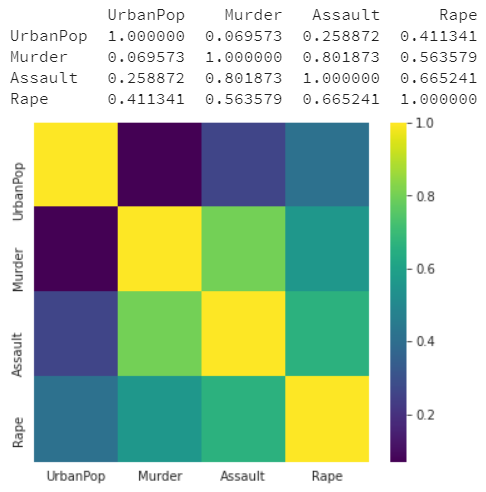
Figure



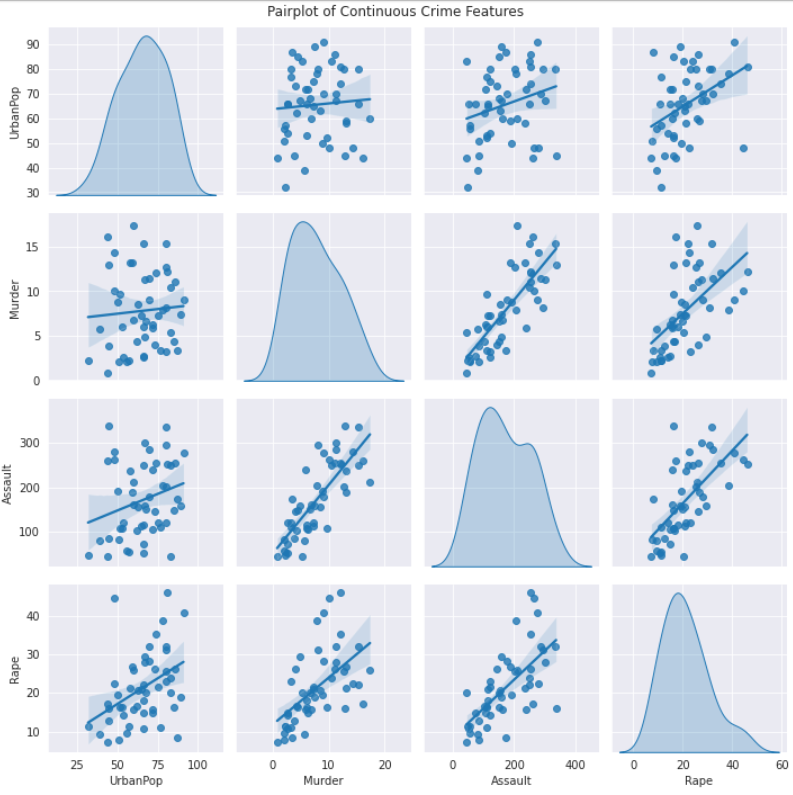
Figure



Figure



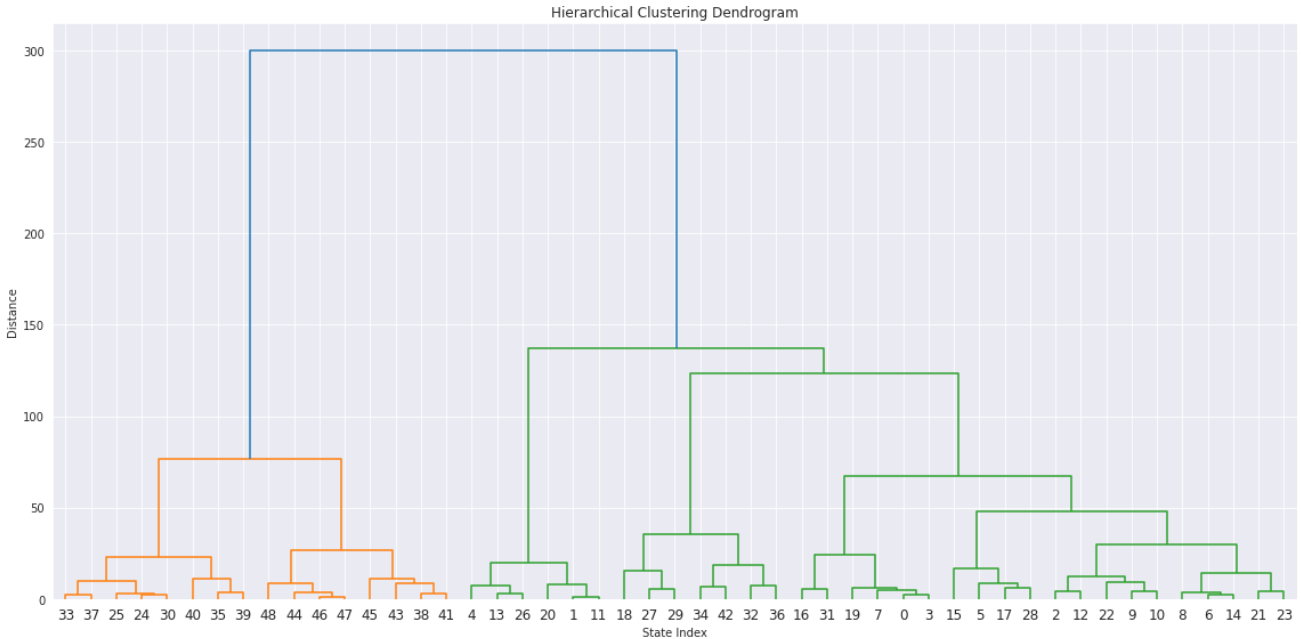
Figure



Figure



Figure



Figure

