Teresa M. Vail

COMP 4449

4/28/23

Midterm Project – Violent Crime Rates by US State

Crimes affect everybody. Beyond the trauma to victims, crimes can indirectly affect the residents in an area by influencing the security measures taken in the area, the amount of people who may choose to move to the area, the sense of well-being by the residents, and far more. Many researchers have taken on the mammoth task of teasing out the influences behind the crime rates in cities, states, and nations to inform policy decisions. I am far from a well-seasoned researcher (nor do I have the social science background), but I have taken a keen interest in crime rates through my data science journey at the University of Denver. In Python Software Development course, my partner and I investigated and visualized the city of Baltimore crimes in correlation to 911 calls. In Data Science Tools 2, another partner I took a deep dive in the mass shootings in the US. It was a natural progression to identify hierarchies of clusters of US states according to violent crime rates data, with a deeper dive on impact of geographic location.

The dataset utilized in the project contains statistics on arrests per 100,000 residents for assault, murder, and rape for each of the 50 US states in 1973. Please note that I misinterpret the numbers of arrests as *number of victims* for my presentation; this has been fixed in my code since then. The numbers of arrests may not accurately reflect the actual numbers of crimes committed – particularly for rape, where it’s difficult to arrest people due to lack of evidence and DNA testing back in 1973, and rape victims may be afraid to call police after the crime was committed. The percent of population living in urban areas is also given for each state. This means that this dataset is initially very tiny – just 50 rows and five columns (State, Murder, Assault, Rape, and UrbanPop).

Since the dataset did not come with any geographic location reference, I added four new columns: Region, RegionCode, Division, and DivisionCode. The four new columns are based on the Census Regions and Divisions from the US Census Bureau. This is the most used classification system according to Wikipedia. The RegionCode and DivisionCode numbers are relative to each other. For example, Northeast region is 1 and will increase as you head westward to the West region, which is 4. This relativity allows me to scale these columns since they would be considered more of a numeric rather than ordinal.

The newly modified data frame has three categorical and six numeric columns. The distribution of numeric columns wide ranging. The Urban Population feature has a central tendency in the center (at around 65), but there is a smaller second peak at 80. In other word, it doesn’t quite fit the Gaussian distribution. The assault, rape and murder features appear to skew right. The region and division distributions is very chunky; region skews left and has a peak at 3, and whereas division may skew left but it’s complicated by four tall peaks all over the graph. The features murder, urban population, rape and division do NOT group around a given tendency, whereas assault and region do.

I used the describe() function to see the mean, min, max, etc. of each numeric feature. It appears that there are far more arrests for assaults for rapes and murders, and all crime rates are greater than the numbers of regions and divisions. Given that most features’ distributions are skewed, I opted to normalize instead of standardizing. Hierarchical clustering needs the data to be on the same scale; otherwise, the results will not be as accurate. After normalizing the data, the assault’s distribution has skewed strongly to left, and the scatterplot shows a strong correlation between the Urban Population Percentage and Assault Arrest Rate. The other two crime’s scatter plots showed the same “no correlation” after normalization of data.

Since I am so driven by the question “Is geographic location driving the crime rates, or is it the urban population percentage?”, I opted to run multiple regression models (OLS) on all crime rates. They are grouped into two categories: divisions and regions. Since regions and divisions are essentially indicating the geographic location but on different scale, it doesn’t make sense to include both regions and divisions in the same model. When I included the regions, the geographic location has a stronger influence than the urban population percentage on all crime rates. When looking at the models with the divisions, the results shifted. While the geographic location still has a stronger influence on rape arrest rate, urban population percentage has equal pull as the geographic location for assault arrest rate, and the urban population percentage have a bigger impact on the murder arrest rate.

To determine how many latent groups there are in hierarchical clustering while still using the lenses of “is geographic location influencing crime rates?”, I conducted agglomerative hierarchical clustering for dendrograms with no geographic locations, with just regions, with just divisions, and with both regions and divisions. In all dendrograms, three clear clusters have showed up, and the subsequent scatter plots for all three crime arrest rates attested to the three clusters. The first two clusters have a tiny bit of intermingling (as indicated by not-as-sharp border), whereas the third cluster is distinctive on each scatterplot. To compare the results of clusters with no geographic location to other dendrograms with various levels of geographic locations (and to each other), I used V-measure. V-measure is the harmonic mean between homogeneity and completeness, and thankfully for us, the metric is symmetric. This means that switching true label (e.g. y\_true) with pred label (e.g. y\_pred) will return the same score value. This is super helpful for us since the real ground truth is not known.

The V-measure score between the no geo location clusters and the region clusters is 1, which means both arrays are the same and it is both homogeneous and complete. The V-measure score between the no geo location clusters and division clusters is 0.916. I was able to trace this to Wyoming switching clusters. However, the V-measure score is still very high, and thus the arrays are very similar. Finally, the V-measure score between no geo location clusters and clusters with both region and divisions is 0.916, which is the same as the V-measure score between the no geo location clusters and division clusters.

In conclusion, geographic location may have a slight impact on crime rates, but it ultimately depends on HOW you’re analyzing the data. If you are using the multiple regression models, then it appears that geographic locations have a stronger influence on crime rate in states, but only consistently on regional level. Additionally, there is a strong R-2 value for assault rate but it got weaker with rape and even weaker with murder. If you look at the multiple regression models at the division level (which is more granular than region), then the data is not as clear-cut. If you are using the hierarchical clusters, there are consistently three clusters with very similar V-measure scores even when playing around with the omission and level of details for the geographical location. This indicates that the geographic location had a very little impact on hierarchical clusters of states. Due to wrestling with the geographic location question, I spent about 30% on data preparation, 40% on data analysis, and 30% on data visualization.