BIG DATA MANAGEMENT

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

DEPARTMENT OF INFORMATICS AND COMPUTER ENGINEERING

ANALYSIS OF BIG DATA AND PRESENTATION OF RESULTS

TEAM DETAILS

DETAILS OF STUDENT 1: ATHANASIOU VASILEIOS EVANGELOS (UNIWA-19390005)
DETAILS OF STUDENT 2 : TATSIS PANTELIS (UNIWA-20390226)
DETAILS OF STUDENT 3: PETROPOULOS PANAGIOTIS (UNIWA-20390188)

PROFESSOR: ANDRITSOS PERIKLIS
SUBMISSION DATE: 06/21/2024
DEADLINE: 21/06/2024

BIG DATA MANAGEMENT

CONTENTS

BIG DATA MANAGEMENT

Introduction

Introduction

In the following work we attempted to compare 2 datasets for common years (2015-2016), with the aim of analyzing and investigating a possible relationship between unemployment and incidents of police killings in different US states. The first dataset concerns unemployment rates , while the second dataset concerns murders by police officers .

We used each state as the common denominator of the 2 datasets, while we then considered it necessary to use an additional dataset that would include the population of cities in the USA . This was done to provide a more comprehensive assessment of our data.

The technique we used to analyze our data was that of clustering and more specifically we used the k-means algorithm. By using this technique we were able to group our data into clusters and observe patterns and correlations within them. To evaluate the quality of these clusters, we relied on non-predictive means such as SSE (sum of squared error) and silhouette coefficient, thus ensuring better quality clusters.

The paper as a whole contains a thorough analysis between our 2 main variables (unemployment, homicide rates and net numbers), contributing to the debate on whether socio-economic factors such as unemployment may be related to the rise in police brutality.

1. Problem definition and motivation

1.1 Description of the Problem

There are many occupations that could benefit from a survey that could correlate unemployment rates by state with corresponding homicides in the same years. Professions that have to do with the study of human behavior (sociologists, psychologists, economists, etc.), can use any findings in other research or delve into the real causes of the problem.

1.2 Research Utilization Incentives

The reasons that may lead to the utilization of the research we do (in the event that it shows some correlation between the data we compare), are many. However, we must not forget that the correlation between 2 factors does not in any way imply causation behind some human behavior. It may be that the factors that lead to murders are other than those of unemployment. Correlation is simply the beginning of understanding possible causes of a problem. However, even this can lead to measures being taken that will provide an improvement in public safety.

1.2.1 Understanding crime

One of the main motivations for someone to look for an investigation that concerns the possible correlation between high unemployment and high incidents of murder, is in the event that some study of crime is being done and some of its causes are being sought. In the event that the reasoning of the investigation is correct and is identified with other similar investigations, then it can be used as a cause of criminality.

We remind you, however, that this does not mean that the problem of crime can be fully understood, as incidents of delinquency are always the result of human behavior.

1.2.2 Creation of social policy actions

Another motivation for verifying a research that relates the economy to a human criminal action is to shape political decisions. If it is shown through research that unemployment can significantly affect the number of murders that take place, there is an additional reason for states to take care of the issue and seek a solution to the problem of unemployment.

2. Brief description of the data set

2.1 The Work Data

The data for the work was drawn from Kaggle, a tool used by data scientists. Our data includes percentages and numbers of people in poverty and the number of police killings in different US states for different years. Specifically, the dataset includes the variables Poverty Rate . Numbers in poverty and Each homicide separately as a registration for each state. The data were used to investigate the possible correlation between unemployment and crime during the period in question. As a key to join these 2 variables we used the state.

During the work we considered it necessary to introduce an additional variable, the Total Population of Cities in the United States and convert it into population per state. The specific dataset was extracted from a well-known data management and distribution platform, opendatasoft.

2.1.1 Data Analysis of Killings by Police Violence

The data we retrieved had to be processed in such a way that it was limited between the years 2015-2016. Their storage for the necessary pre-processing is done in .csv files. From these years we had data for the whole year. Some of the information in the table of killings by police in the years we mentioned are: the state where the killing took place, the date, the number of bullets used, the cause of death, the age of the victim, the city where the killing took place .

There are other disturbing insights in our data, which are overwhelmingly similar across the 2 years. All the victims were men, almost all had no signs of mental illness, and almost none of the murders involved the use of body cameras.

Our final data was limited to the State and the number and percentage of deaths in it.

2.1.2 Analysis of Unemployment Data by State

The unemployment data by state is also limited to the years 2015-2016, because we want to deal with these specific time periods. These data were also saved in .csv files so that we could process them properly. Some of our initial data items were the state, its number of unemployed, the year, and the percentage of unemployed relative to the state's population.

Our final data were limited to state, number of unemployed, and unemployment rate.

2.1.3 Analysis of Population data by State

Population data were subsequently introduced into the work as a measure to confirm the results. Their content concerned the year 2015. The population was for American cities with a population greater than 65,000 people. The processing of this data was decisive for the final conclusion of the paper, as it gave us a better connection between the total population and the results we found from using the 2 previous datasets.

3. Description of the data analysis method

Before we start the analysis at this point, let us recall that the dataset with the population was included at the end more experimentally to confirm the results of the experimental part.

3.1 Data Preprocessing

In this subsection, we will analyze all the preparation that had to be done on the data before clustering.

3.1.1 Data Pruning

As we analyzed in the previous chapter, we use two datasets where each has a different time range (specifically the dataset with murders has data from 2015 to mid-2017 and the dataset with poverty rates for the years 2011 to 2021). The data we will use must coincide in time, otherwise there would be no point in trying to correlate them. Therefore, we chose that the data should be limited to the years 2015, 2016 and not the year 2017 because the homicide data was not complete until 31/07/2017 only.

3.1.2 Defining a union element of datasets

Since we have 3 different datasets, we had to find their "connecting link", i.e. the element in which the two datasets will be joined. Recalling our primary goal, which is to prove that states with high rates of unemployment also have high incidents of murder - correspondingly for low levels of unemployment we have few incidents of murder - we concluded that the join should be done on a state-by-state basis.

3.1.3 Transform data to a common scale

One of the most important problems we have been asked to solve is that of finding and transforming data at a common scale. Initially the poverty data contained the percentage of unemployment for each state ("Percentage in Poverty") and numbers of unemployed. However, in the homicide dataset, it did not contain the number of homicides per state and apparently based on this neither did the homicide rates per state. So our job was to first count the murders per year and the murders in each state in "net" numbers and then convert them into percentages for each state. Also in the dataset with the population, again we did not have the population of each state, but the population per city in America (for a population > 65,000 inhabitants). Therefore we had to add up the population from each city of each state for this as well.

3.1.4 Transforming data into common morphology

An important problem was also the different name with which the states were listed in the data we had found. In the homicide dataset the states were written in abbreviation (eg NW for New York), while in the poverty rate dataset they were written by their full name. So we had to clean up a common morphology. We decided that we should convert the full names of the poverty dataset into an abbreviation with the help of a mapping table.

Finally, for formal reasons in addition to the calculation of the murder rate, in the remaining fields (age, sex, race, etc.) we selected a representative value in each state based on the most frequent occurrence of murder incidents.

3.1.5 Kill Sum

As we said the homicide data was in registration form, meaning each registration was a separate homicide, so we counted each homicide separately for each state and also calculated the percentage relative to total homicides.

3.1.6 Fixing no-kill states

Some states had no murders for a given year (eg Rhode Island had 0 murders for the year 2015) so this should be taken into account when joining the tables, otherwise it could not be included (enter a value of 0) .

3.1.7 Correction of data gaps

The data was very good and there were no gaps

3.1.8 Normalize data as standard deviations

Data before entering the algorithm were transformed to standard deviations around 0

3.1.9 Logarithmic normalization

The raw data, i.e. the absolute number of unemployed in each state and the absolute number of murders in each state before applying the algorithm, were normalized using lognormalization due to the difference in volume (due to the difference in population).

3.2 Data Analysis

For data analysis there are many different techniques that can achieve this goal (eg clustering analysis, regression analysis, factor analysis). In this work we used cluster analysis.

3.2.1 Why Cluster Analysis

- It helps to better understand the differences of the groups and draw their conclusions.
  - In our case, the understanding of why some American state may have more murders than another.
- Easy to implement.
- It can provide quick conclusions without interfering too much with the data.
- Detection of unusual data, that is, if some data are outliers, they may fit themselves into a small cluster, which is a sign that they are exceptions.
- It does not require data tags.
-

3.2.2 Why K-Means

- Faster than the other categories of clustering such as hierarchical and densification
- Better on not so big data (51 in our case)
- It is simple and easy to implement
- Good solution in our case where we don't have data whose labels we know, as for example we have in the iris data
- It is good at detecting hidden patterns especially in data where we don't know from the start its spatial footprint.

3.4 Cluster Quality

3.4.1 Sum of Squared Error

Generally SSE is a measure that calculates the deviation of the actual values from the predicted values in a data set

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

It can in clustering be used as a measure to understand how close the data is overall across all clusters with formula:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

3.4.2 Contour Factor

In a clustering it is very important that the data in each cluster are very close to each other (cohesion) and well separated with respect to the other clusters (separation). The contour coefficient is a measure that takes both into account.

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

Where a is the average distance of i from the points of its cluster and b is the shortest distance of i from the nearest point of another cluster. It takes values from -1 to 1. A positive value means that the clustering is acceptable (the more closer to 1 the better) while a negative value is not acceptable.

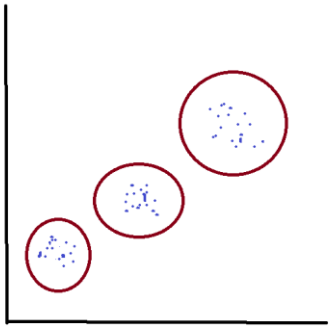
3.2.5 Analysis Stages

1. In the first stage we used unemployment rates by state and homicide rates relative to total homicides by state. The results of this phase and after reviewing the way of analysis we decided that we had to change our methodology.
2. In the second stage we used the pure numbers. That is, the absolute number of murders per state and the absolute number of people in poverty per state.
3. In the Third stage we used logarithmic normalization of the absolute numbers.
4. Finally we used the third dataset with the population to normalize the data to pure numbers.

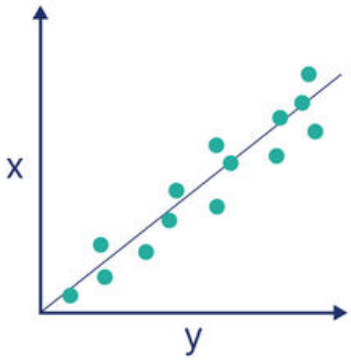
The reasons we changed the methodology and went from the first phase to the last will be explained in the critical discussion of the results.

3.5 Results We Expect

At this point we will show how we imagine the results to draw the conclusion that there is a correlation



or



4. Experimental results

4.1 Percentage results for the years 2015, 2016 and 2015-2016 average

4.1.1 Percentage results for the year 2015

Poverty rates and homicide rates

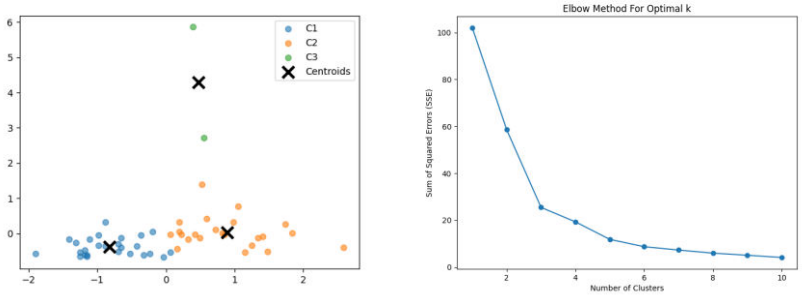


Figure 1

SSE (Sum of Squared Errors) = 25,464

Silhouette Coefficient: 0.518

4.1.2 Percentage results for the year 2016

Poverty rates and homicide rates

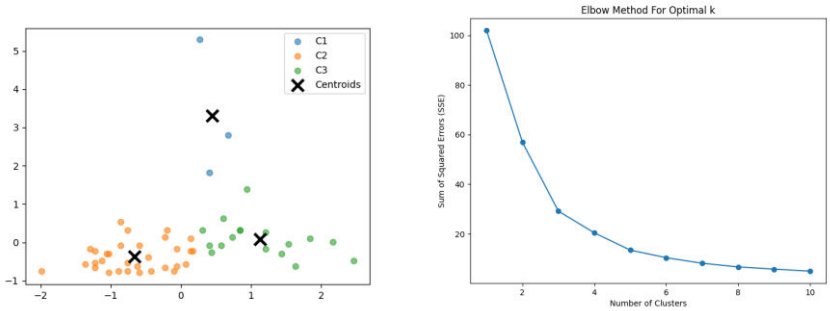


Figure 2

SSE (Sum of Squared Errors) = 29,151

Silhouette Coefficient: 0.495

4.1.3 Results of average percentages for the years 2015-2016

Poverty rates and homicide rates

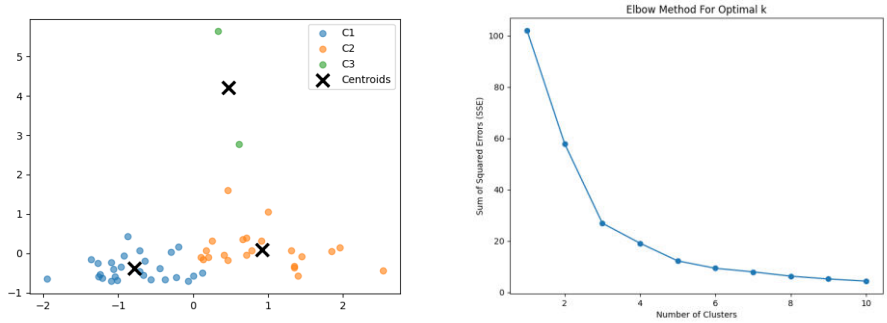


Figure 3

SSE (Sum of Squared Errors) = 26,879

Silhouette Coefficient: 0.501

4.2.1 Results of net numbers for the year 2015

Number of people in poverty and number of murders

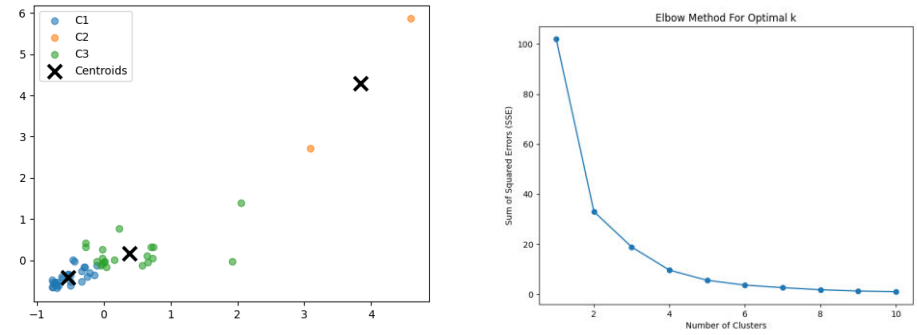


Figure 4

SSE (Sum of Squared Errors) = 18,885

Silhouette Coefficient: 0.481

4.2.2 Results of net numbers for the year 2016

Number of people in poverty and number of murders

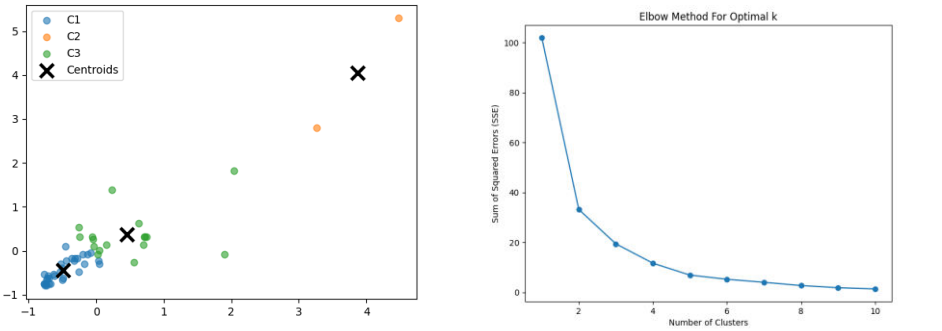


Figure 5

SSE (Sum of Squared Errors) = 19,478

Silhouette Coefficient: 0.484

4.2.3 Results of net numbers log normalized for the year 2015

Number of people in poverty and number of murders

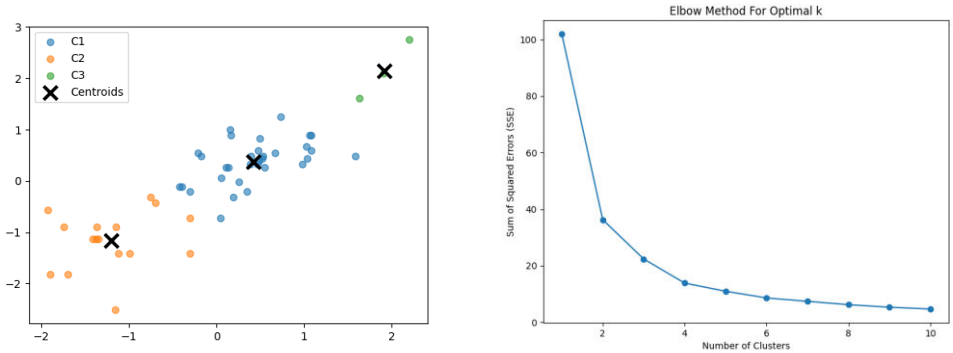


Figure 6

SSE (Sum of Squared Errors) = 22,391

Silhouette Coefficient: 0.554

4.2.4 Results of net numbers log normalized for the year 2016

Number of people in poverty and number of murders

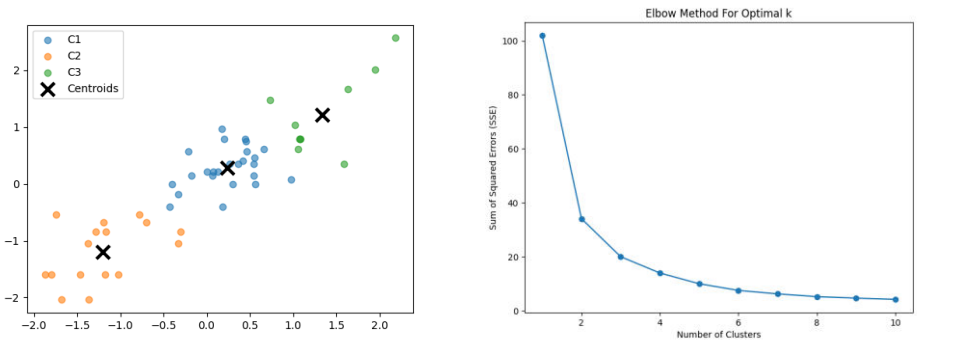


Figure 7

SSE (Sum of Squared Errors) = 20,014

Silhouette Coefficient: 0.491

4.2.4 Results of net numbers normalized with population for the year 2015

Number of people in poverty and number of murders and number of population.

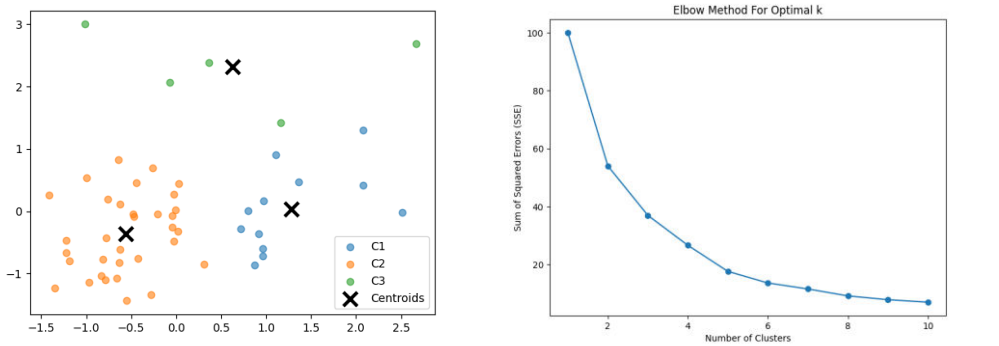


Figure 8

SSE (Sum of Squared Errors) = 36,972

Silhouette Coefficient: 0.453

5. Discussion/Critical evaluation of results

In the data we will see, each point is a state. The x and y axes change based on the most field of the dataset we choose. The x axis belongs to the category of police violence and the y to unemployment. The number of k in the pre-clusters is determined based on the elbow method.

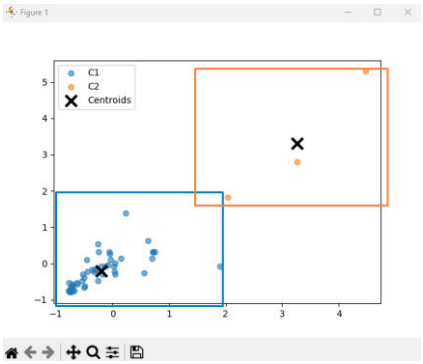
5.1 Critical Analysis of Rate Results for the Years 2015, 2016 and 2015-2016 Average

For the years 2015, 2016 and a combination of these using an average (images 1, 2, 3 respectively) we notice that the conclusion we want to prove, that is that in the states with more unemployment there is more police violence is not reflected since the rate of murders does not increase as unemployment rates are rising. We notice that the graph generally has a norm of points below y=1. In short, as murders increase, so does unemployment.

The first phase of the analysis had disappointing results. But at this point we realized that we have made an error in the analysis. We have not included the population factor which is very important. It makes sense that states with large population rates (such as California) would have higher homicide rates and smaller states (such as Rhode Island) would have lower rates. The logic with the subsistence rates was also incorrect, the poverty rate given to us by the dataset has been calculated based on the population of each state (the population data is not given to us and this is one reason we use our own dataset with population below) while the percentage of murders per state (which we calculate) has been calculated based on the total number of murders and not on the population. So we compared results with a different denominator. The dataset also had net numbers of people unemployed and we have calculated the corresponding murders in each state. We therefore used these data for further analysis.

5.2 Critical Analysis of Net Results for the years 2015, 2016

After applying K-Means to the data (figure 4.5) we notice a more promising picture. Most of the data for both years are clustered before x = 1 and y = 1. So we see a norm at this point, while there are some data that are very far from it.



If we exclude the second cluster where it seems that the data do not follow the norm, we see in the data of the first that there is generally an increasing trend, a result that satisfies the conclusion we want to draw. However, here again there is an error in the analysis. Again we do not have the data normalized in relation to the population that we mentioned in 5.1

5.3 Critical Analysis of Results of logarithmically normalized net numbers for the years 2015, 2016

The data in images 6 and 7 provide us with a very good result. We see a clear upward trend and clusters are relatively good by both measures. With this image we can say that there is indeed a correlation between these two elements. But because log normalization is blind. That is, it does not know the population but simply applies a mathematical formula that compresses data with unequal variance, the ideal would be to confirm it through the real population per state. This is exactly what we will do in the next step (here it should be noted that the population data refer only to the year 2015 because we could not find data for 2016).

5.4 Critical Analysis of Population-Normalized Results of Net Numbers for the Years 2015, 2016

In figure 8 we see a not so good picture as in that of the logarithmic normalization. We notice that there is again an increasing trend, however the data is quite sparse (it is reflected in both the sse and the contour coefficient) so that we cannot safely conclude that there is a correlation. However, it is the most correct methodology of all because it responds to real population data and not to blind methods and assumptions. The error of the method is that the population refers to cities that have a population >= 65,000 inhabitants not counting the smaller cities.

5.5 For future analysis

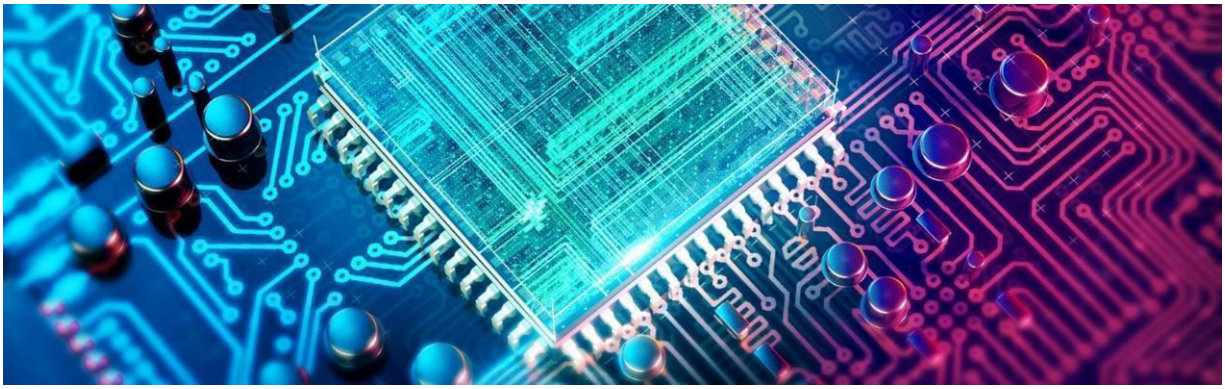
Some things that could be done for more rational results would be to express the population data in terms of homicides and people in poverty as percentages of 100. This might bring the data closer. Also, a better dataset could be used with a population that includes all cities in the states and not just the largest.

Conclusions

Social data is complex and multifactorial. A two-trait analysis is not necessarily able to prove a link between police brutality and unemployment. For safer results, other dimensions should be introduced in the analysis. But for the specific data we have and the reasoning we followed due to the incremental nature of the data more in the log normalization less in the normalization based on the population and the pure data (we do not count the initial assumption we made in 5.1 because it is based on wrong bases.) we believe that there is a correlation, but not so strong that we can confidently draw the conclusion that it is true. In short we would say that there is an indication based on the data that requires verification from other data as well.

BIG DATA MANAGEMENT

Thank you for your attention.



BIG DATA MANAGEMENT