Université Claude Bernard Lyon 1
Institut de Science Financière et d'Assurances

STUDY RESEARCH

Criminality and its factors, an Econometric analysis

Members
PARCIGNEAU Victor LOGEL DEMOULIN Suzelie

Referring Professor
HAVET Nathalie

2022-2023

Abstract

In this study, we analysed crime data from 82 different countries. We then examined the variables that we thought had the greatest impact on crime, such as socio-economic factors (unemployment, average wage, resources allocated to public services, etc.), but also more controversial variables, such as immigration, legislation or religion.

With this study, we wanted to validate or refute certain socio-economic theories regularly invoked to justify reforms and political orientations. To this end, we created three crime indexes, based on justice statistics, to characterise delinquency in each country.

We then used this index in an econometric model to find out what the best responses are when a nation wants to reduce its crime. With these results, we hope to be able to better analyse and criticize domestic policies and their effectiveness.

Sommaire

Introduction

Crime can be defined as all illegal, delinquent and criminal acts. It is a complex and multifaceted phenomenon, ranging from simple theft to murder and sexual violence. It is characterised by the transgression of the legal norms of a social system. One of its objectives is to maintain order in society by demanding a certain behaviour from its members. Criminal norms are regulatory and coercive, they form the legal order and thus contribute to the peacekeeping.

Criminologists are of the opinion that these norms are socially enacted and therefore vary from country to country, which means that crime as universal does not exist. This was a first difficulty, but we will come back to this point in more detail later.

Some behaviours have been labelled as 'crime' and have subsequently been prohibited. But other criminologists think that norms are objective, because they can be found everywhere and always, such as killing someone without reason. But "without reason" can differ greatly in different societies and periods. Therefore it can be quite difficult to define crime in a comprehensive way and we humbly propose a definition that will be presented and argued later.

The study of crime is a complex and controversial subject that has attracted increasing interest in recent years. Crime has a profound impact on the lives of citizens, public policy and society as a whole. Understanding the causes and consequences of crime is therefore crucial for developing effective crime prevention and control policies.

Over the years, researchers have studied crime from different perspectives, seeking to understand its causes and consequences, as well as ways to prevent and fight it. However, crime remains a persistent problem in many countries, and its impact on communities cannot be ignored. Crime levels vary considerably from country to country, and even within countries. We believe that factors such as the level of economic development, social conditions and crime control policies in place are the main factors that influence crime levels.

This study is also motivated by the rise of populism in many countries around the world. Indeed, this political trend has often been associated with growing concerns about security and crime, as well as distrust of political institutions and elites. In some cases, populist governments have sought to strengthen anti-crime laws and policies, while in other cases they have adopted more controversial policies, such as cracking down on minorities or limiting civil liberties.

In this context, we believe it is important to continue researching crime and ways to prevent and fight it. By examining the factors that influence crime, we can better understand the underlying causes of the phenomenon and propose effective solutions to deal with it.To carry out this research, we will mainly use data analysis and econometrics. Indeed, these two subjects form the core of our Master's programme and represent our future tasks quite well in our opinion.

In this research, we will look at the analysis of crime, examining the different factors that can influence it,weither they are socio-economic, cultural, political or other. We will also look at the prevention and repression measures put in place by the authorities and their effectiveness. In sum, the aim of this work is to contribute to a better understanding of crime and to suggest ways to improve public policies in this crucial area.

Chapter 1

Crime Indexes

In this chapter, we will further analyse the concept of crime. Through this research, we will present the index that we have created for this work and the variables used in it. We will compare it with the different indices that already exist and explain its strengths. Then we will look in more detail at the creation of our index and its limitations.

The crime index is essential to this study, as it is the variable to be explained in our econometric model. It impacts the entirety of our work and its conclusions, which is why we must be very rigorous in its creation.

1.1 What is Crime ?

1.1.1 For the sociologists and criminologists

When we talk about crime, we regularly make the mistake of talking about the criminal and not the crime. The most famous example of this way of thinking is the work of Lombroso and the italian positivist movement of the early 20th century. In his book "The Criminal Man"[15], this Italian physician defends the innate character of certain criminal behaviours.

This type of thinking was widespread, particularly because of the racist theories that were widespread at that time. It was thought that a human being was predestined to become what he or she would become, that it was written in the genes. With this type of thinking, it is easy to conclude that crime is due to certain categories of the population and that they are entirely responsible for it.

After the Second World War and its atrocities, we realised that this kind of thinking, apart from being extremely dangerous, was absolutely wrong. The human behavour in a community can't be summarized and justified only by his biological characteristics. Before sociology started to analyse crime, the individual was considered apart from his group, outside of any context. This obscures the fact that "man is a social animal", according to Aristotle.

In his article "What is crime?"[18], Gabriel Tarde denounces this thinking mechanism and argues that crime is a social construction. He concludes that social norms and community expectations are key factors in defining what is considered criminal.According to him, crime is not an objective entity that exists outside of social processes, but rather a social construct that is the result of the interaction between social norms, individual perceptions and concrete actions. He suggests that the perception of crime varies according to social, cultural and historical contexts, and that the definition of crime is therefore constantly changing.

Etienne de Greef explains, in his book "introduction to criminology"[19], that after the failure of Italian positivism, sociologists turned to a societal analysis: "It thus becomes quite possible to explain an offender's career, solely as a result of the circumstances in which he or she found himself or herself."(De Greef, 1946, p37)

However, he later added that this analysis was of little use when it came to examining the criminal: "Faced with a criminal who committed his crime in June, we can certainly say to ourselves that he is linked to the general curve given to us by statistics and has therefore obeyed some law. But since not everyone becomes a criminal, even in June, what we need to find is the reason why he has found himself sensitive to this factor."(De Greef, 1946, p37)

This is why in this study we will not look at crime at the level of the individual, as we could have done with a probit-logit model for example, but at the level of a state. We want to make explicit the laws that this statistic follows, in order to help policy makers take the right measures to reduce this phenomenon.

Now that we have justified the choice of scale for the analysis of this study, let us look at the definition of crime itself.

1.1.2 With regard to the law

First of all, when we talk about crime, we include both delinquents and criminals. We therefore consider everything that is considered illegal by a state.In the following, we will take the example of French criminal law[9]. In this law, there are three types of criminal offences: contraventions, délits and crimes.

"Contraventions are offences which the law punishes with a fine not exceeding 3,000 euros" art. 131-13

"The law shall determine the crimes and offences and set the penalties applicable to their perpetrators. The regulations determine the contraventions and set, within the limits and according to the distinctions established by the law, the penalties applicable to offenders." art. 111-2

However, the law is bound to change over time first and according to the country even more so. Therefore, criminality as an immutable and universal thing does not exist. However, this does not mean that it is impossible to estimate this data at a given moment.

It is true that in our previous analysis we presented crime as a term that is difficult to define, especially in practice. However, this does not mean that there are no legal definitions of crime.

Legal crime refers to offences defined by law and sanctioned by the state. This includes crimes and offences such as theft, assault, murder, fraud, etc.

Apparent crime, on the other hand, refers to criminal acts that are reported to the authorities. It can therefore include crimes and offences that are committed but also reported to the relevant authorities.

Finally, actual crime refers to the totality of crimes committed, whether or not they are reported. It is therefore a broader measure of crime that takes into account all delinquent behaviour, including that which is not reported to the authorities.

In this study, we will limit ourselves to measuring apparent crime, as this is the only one for which we have statistics.

1.1.3 Scientific Literature

At the beginning of our study, we used the crime index offered by Numbeo[3], which is freely available. Numbeo is a global database, based in Serbia, on local life prices, crime rates, quality of health care, among other statistics.

Their data collection is based on collective intelligence, including the use of surveys. This method is very effective for economic types of data, such as the price of living or real estate. However, when it comes to more controversial data, it is more complicated. Indeed, it seems to us to be rather flawed to measure crime through the feelings of the inhabitants.

First of all, it is biased because only the large cities are represented and aggregated to form the country's score, which obscures the rest of the territory. Secondly, some crimes are probably disproportionate to their real importance. For example, a murder or a burglary are spectacular crimes, with a great impact on the feeling of insecurity. Finally, this index can very easily be manipulated by third parties, as shown by the fact that recently, Nantes and Brest had temporarily become the most dangerous cities in the world, which the extreme right had used to denounce the "impoverishment of the French nation".

In the end, this index is a fairly accurate representation of the feeling of insecurity, but it is of little interest to us here. We want to base ourselves on statistical facts, not on statistics. However, to our great surprise, no one else has taken the trouble to create such an index. We will therefore adress this shortcoming in the rest of this work.

In his study, Crime and Unemployment in Ireland, 2003-2016, Enda Patrick Hargaden examines the relationship between the unemployment rate and crime in Ireland between 2003 and 2016. The author uses Irish police data to examine crime trends over a 14-year period, focusing on crimes against persons, property and drugs. He places particular emphasis on the need to weight these values against the population, which we will do on our data. He then used several linear regressions to obtain the impact of the unemployment rate on each crime data.

This method is interesting because it allows us to have a more precise idea of the impact of each of our explanatory variables on the different types of crime, but it is too tedious in our case since we have 13 categories of crime data.

In order to create a new data on crime, Canada created in 2009 the Crime Severity Index (CSI)[12]. The CSI is a measure of crime severity and is based on Canadian police data. It measures the severity of the offence according to two factors: the nature of the offence and the maximum penalty for that offence.

The CSI combines data on all criminal offences reported to the police and weights them according to their seriousness. The most serious offences, such as violent crimes, have a higher weight than less serious offences, such as minor offences. Serious offences also have a higher weight if the maximum sentence is higher.

This method is very interesting because it allows us to combine the different crime statistics into one, losing the minimum of information through weighting. We will use a similar method to calculate one of our crime indices.

The article "The cost of crime to society: new crime-specific estimates for policy and program evaluation"[16] published in 2010 in the journal Drug and Alcohol Dependence aims to estimate the cost of crime to society using a crime-specific approach. The authors (McCollister KE, French MT, Fang H.) used an economic methodology to estimate the direct and indirect costs of crime based on data collected from sources such as police reports, victimisation surveys and health expenditure data.

The results of the study indicate that the cost of crime varies considerably depending on the type of crime. The highest costs are associated with violent crime, followed by drug and property crime. The authors also found that the indirect costs of crime, such as lost productivity due to injuries and deaths, were often greater than direct costs such as health care expenditures.

This method is the most economically consistent and allows us to assign a weight to each type of crime relatively easily. The index calculated from this method is likely to be the most successful and the one with which the analyses will be most interesting.

1.2 How did we create this index ?

1.2.1 Selected variables and sources

In this study, we decided to choose the year 2018, as we needed a year that was close enough to us, but prior to the covid-19 crisis, so as not to bias our data. As for the nature of the data, we based ourselves on the UN definition of crime. Indeed, the UN[2] lists what it considers to be the main illegal acts:

- Use and traffic of various drugs: Cannabis, Cocaine, Amphetamines (including amphetamines and methamphetamines), Ecstasy and its derivatives, Opioids (including prescribed opioids and contraband), Opiates, Tranquilizers and Sedatives (including barbiturates and benzodiazepines) and New Psychoactive Substances or NPS (mainly including ketamine)
- Intentional homicide
- Violent and/or sexual crimes (including acts intended to induce fear or emotional stress, child pornography, kidnapping, robbery, violence against the person, sexual exploitation and all categories of sexual violence)
- Economic crimes (including illegal waste dumping, environmental pollution, burglary, corruption, fraud, money laundering, migrant smuggling and all categories of theft)
- Firearm trafficking (including the number of weapons seized by justice, customs authorities and returned to the police)
- Human trafficking

This list, while not covering all illegal acts, includes many of the most important ones, both in terms of number and seriousness. It also has the advantage of being based on reliable sources and of mentioning acts that are rarely mentioned when talking about crime (notably pollution and white-collar crime)

Moreover, these lists are quite comprehensive and are verified by one of the most reliable institutions in the world: the UN. For the missing data, we relied on data from the states in question to complete our database.

From this, we decided to use 3 different crime indices, in 3 different econometric models. The first one is the Numbeo index, based on insecurity, which we nevertheless find interesting to analyse in order to find the best tools to fight against this phenomenon.

The second will be created by our group and will use the study on the tangible cost of crime that we presented earlier. This will help us to define the weights to be attributed to each variable.

Finally, the last index created will be based on the same method as the Canadians used to design the CSI index. However, as we could not find data on the average sentence for each type of crime, we will use the distribution of the prison population in order to have similar results.
We will now detail the method for constructing these three indices.

1.3 Technical details

1.3.1 Indexing Criminal Data

In order to create these indices, we first had to index each crime variable independently. This was because our table was made up of variables expressed as a percentage or rate per 100,000 inhabitants. They are also extremely disparate, so it was impossible to simply agglomerate them without first transforming them.
To illustrate our method, let us take the example of cannabis consumption by country.

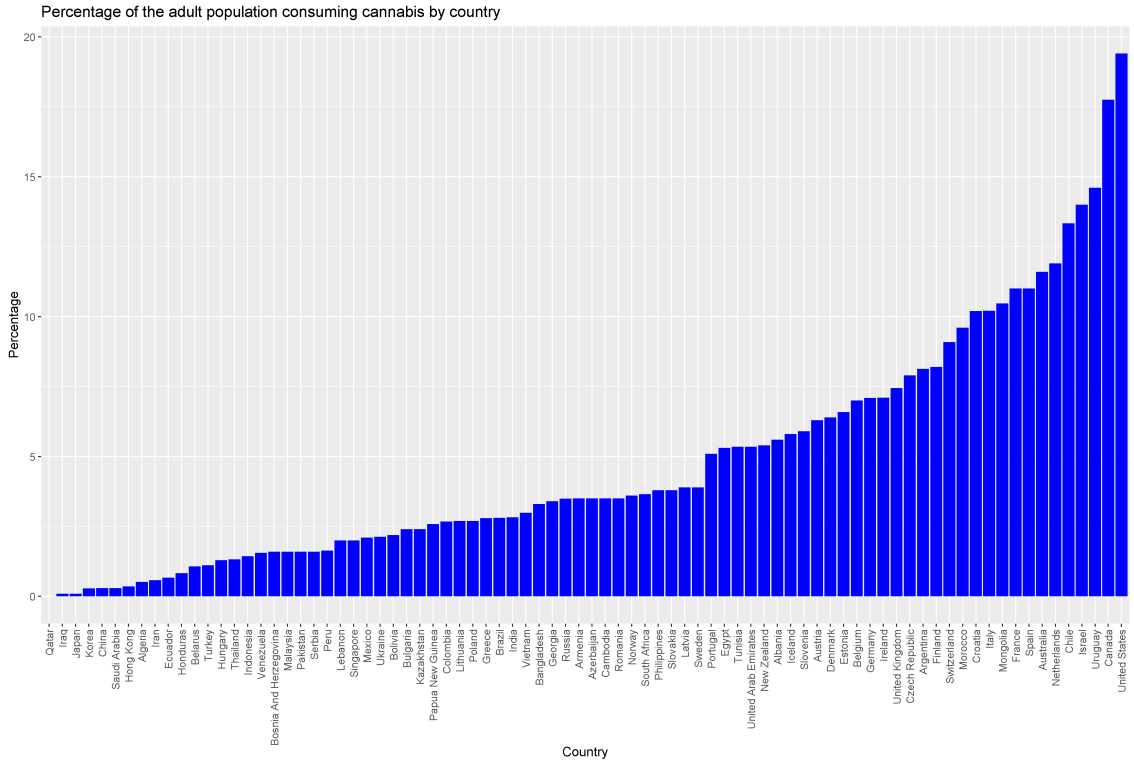

Figure 1.1:

We can see in the graph 1.1 that the distribution is very uneven : the median is 3.5 while the mean is 4.8. The extreme values to the right of the axis have a strong impact on the distribution, despite their small number. Indeed, 80% of the values are below 7, but the values go up to 19.4.

Without these extreme values, we could have used the Max-Min method to transform this percentage into an index. In this method, the smallest value in the sample is associated with the index value 0 and the largest value with the index value 100. For the other values, we proceed in a linear way.

In our case, we cannot use this method because too many values are close to the minimum and maximum and we would lose too much information.
All our data types had the same kind of defects. So we decided to use a non-linear method to create our indices. After a lot of research and unsuccessful attempts, we decided to turn to the sigmoid function.

The sigmoid function is a mathematical function that takes as input any real number and returns a value between 0 and 1. It is defined in its basic form by the following equation:

$$y = \frac{1}{1 + e^{-x}} \tag{1.1}$$

where e is the mathematical constant (about 2.718), x is the input value and y is the output value

The sigmoid function is commonly used in modelling non-linear relationships between variables. In particular, it is the foundation of Gompertz curves[11], which are often used to model the growth of populations, such as human population growth, the growth of bacteria or cancer cells, or the growth of the economic population.

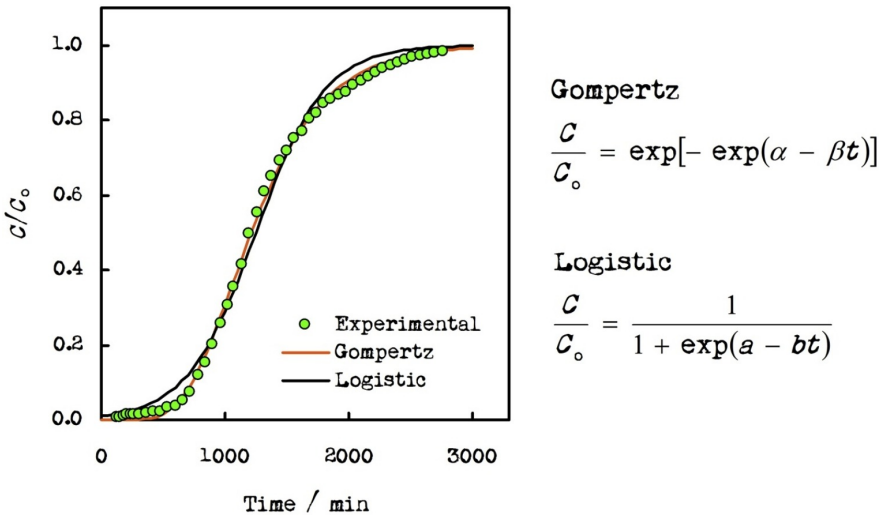

Figure 1.2: Mathematical modelling of solute transport kinetics through a porous medium

It can be seen in the figure 1.2 that the logistic function is a specific form of the sigmoid function, similar to Gompertz curves.

In our case, we will apply this formula to our data:

$$y = \frac{100}{1 + e^{-c(x-d)}} \tag{1.2}$$

where e is the mathematical constant (about 2.718), x is the input value and y is the output value, between 0 and 100, c will vary with each data type to determine the steepness of the curve and d is the median of the x data.

In the case of cannabis use, we chose c = 0.3 and the median is equal to 3.5. We obtain this graph, which represents the values returned by the index and the basic values:

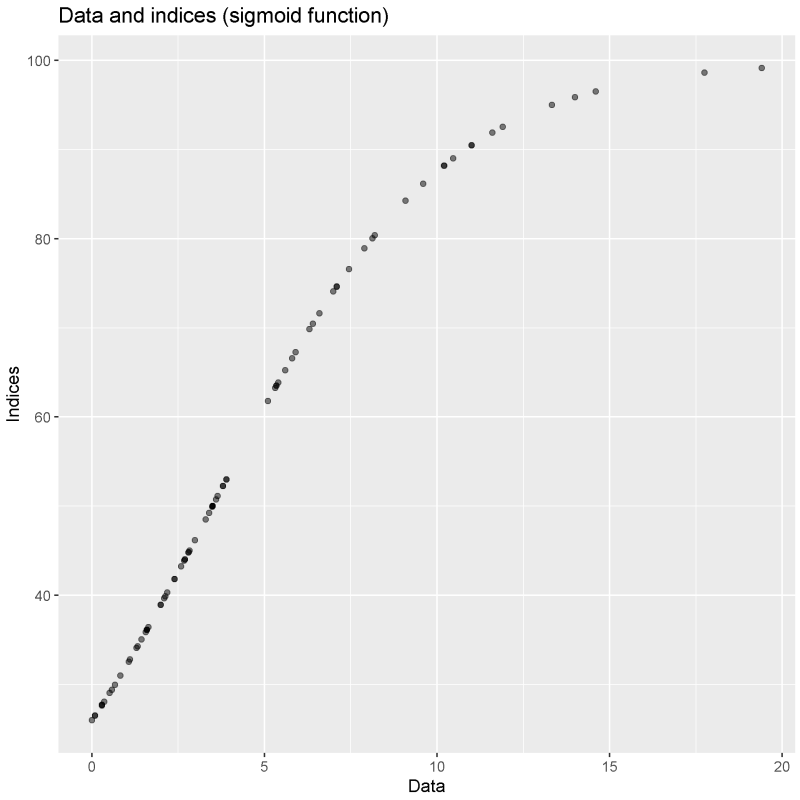

Figure 1.3:

Using this method, we created indices for our 13 categories. Moreover, in both of the weighting method we use, we lack specific data for each kind of illicit substance. With this in mind, we had to create a Drug indices, to reunite all the drugs consumption in one index.

In order to choose the weights associated with the variables, we base ourselves on the study "Social Cost of Drug Consumption"[6] Caulkins, J. P., Pacula, R. L., Paddock, S., Chiesa, J. (2002). In this study, they find those proportions of costs :

Table 1.1: Social Costs of Drug Consumption, as a proportion of the total

Cocaine	**Opiates**	**Cannabis**	**Opoids**	**NPS**	**Ecstasy**	**Tranquilizers**	**Amphetamines**
59.1%	24.9%	8.6%	2.53%	2.47%	1.1%	1.05%	0.25%

With these figures 1.1, we can deduce the following map, representing the drug consumption index per country :

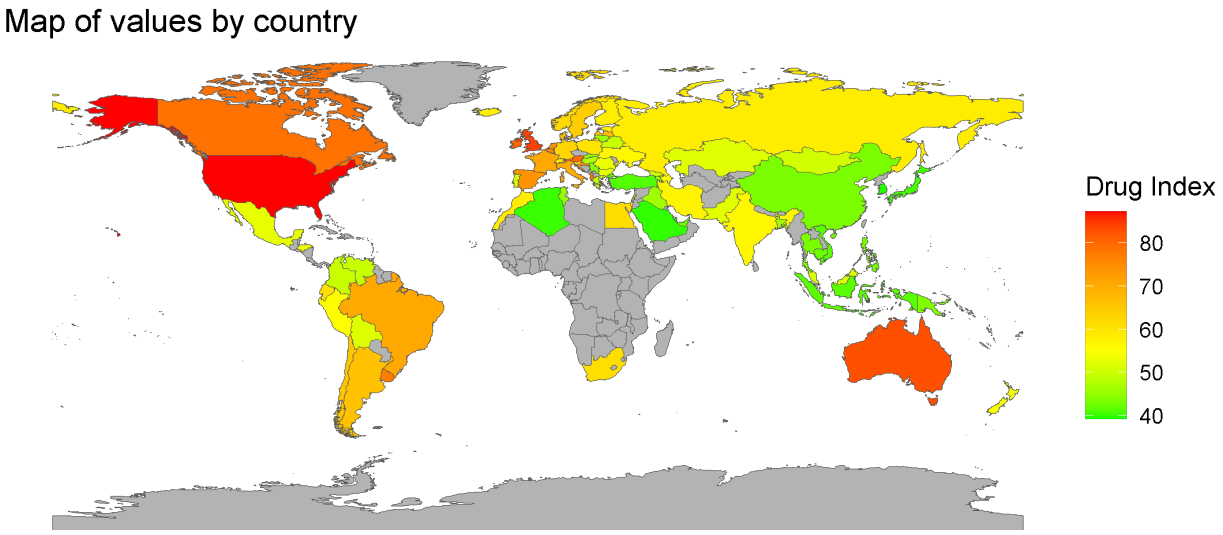

Figure 1.4:

From these values 1.4 and those of the indices associated with the other variables, we can now create our global indexes, which provide a representation of crime.

1.3.2 Weighting using Tangible Cost

In the study "The cost of crime to society: new crime-specific estimates for policy and program evaluation."[16] McCollister KE, French MT, Fang H., researchers try to estimate the costs associated with different forms of crime in the US. The authors used a method based on victim surveys, police and court data to estimate the direct and indirect costs of crime.

They estimated that the total cost of crime in the United States was about $ 1.67 trillion in 2005, or about 6% of GDP. The authors also estimated specific costs for different types of crime, including costs associated with drugs, homicide, sexual assault and robbery.

In our index, we will use the specific costs for each type, which we report in the following table, expressed as a percentage of the total cost:

Table 1.2: Total Costs of each forms of crime, as a proportion of the total

Intentional homicide	**Violent or sexual crimes**	**Corruption and Economic Crimes**	**Human Trafficking**	**Firearm Trafficking**	**Drug Consumption and Trafficking**
63.49%	2.46%	0.85%	3.4%	2.5%	27.3%

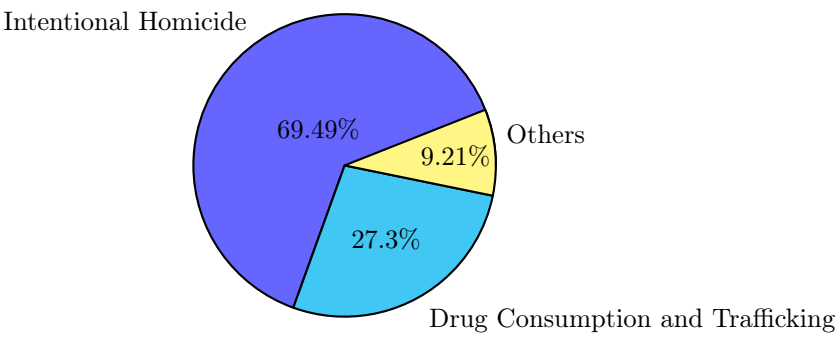

Figure 1.5: Proportions of each form of crime in total costs

Applying these figures 1.2, we obtain the following map, representing crime in 2018 in the selected countries:

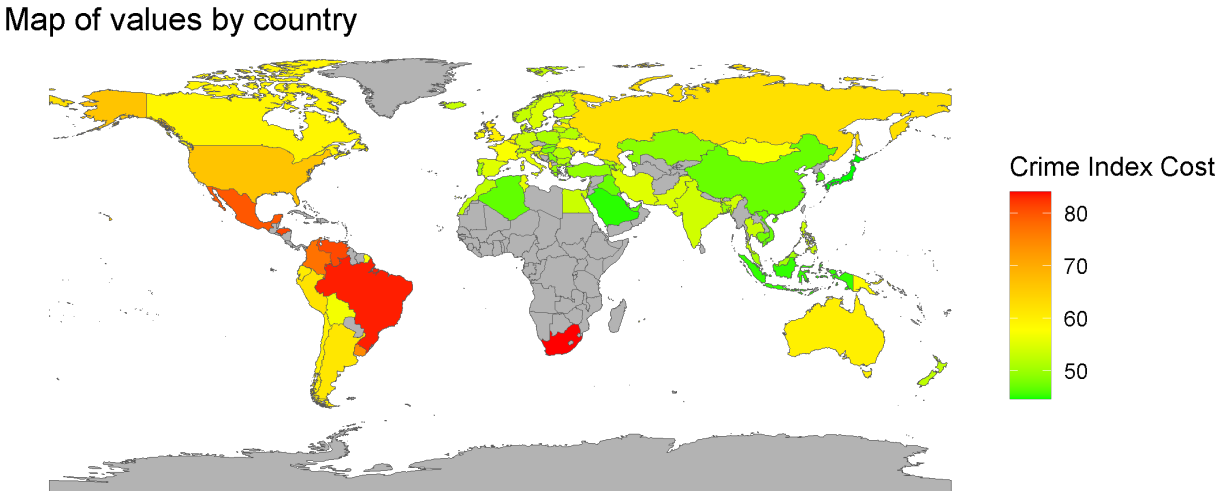

Figure 1.6: Map of the crime index by country, using the tangible cost method

As we can see in figure 1.5, this index is mainly based on data on intentional homicides and drug trafficking. Based on this observation, we thought it would be interesting to create another index, with different proportions.

1.3.3 Weighting using carceral distribution

To create this second index, we based ourselves on the CSI[12] (Crime Severity Index) method mentioned earlier. This takes into account both the frequency and severity of criminal offences.

To calculate the CSI, Statistics Canada uses data from police services in all regions of Canada. Crimes are classified according to their severity, which is determined by the maximum penalty provided by law and the average sentence proclaimed. For example, the most serious crimes, such as murder, have a severity of 10, while less serious crimes, such as simple theft, have a severity of 1.

For each type of crime, the number of crimes committed is multiplied by the severity of that crime to obtain a total severity score for that crime type. The total severity scores for each crime type are then added together to obtain an overall severity score.

The CSI is calculated by dividing the overall severity score by the total number of crimes reported and then multiplying this result by 100 to obtain a numerical index. Thus, the CSI is an indicator of the average severity of reported crime in a given region, which can be compared to other regions or years to observe crime trends.

Unfortunately we could not find data in Open Data concerning the average sentence for each type of crime. Therefore, we had to fall back on a similar data: the occupancy rate of prisons.

Indeed, at a given moment, this figure gives a good idea of the seriousness of each type of crime, since a murderer will stay in prison longer than a thief, all other things being equal.

For this study, we will take the French prison statistics for 2018, considering that they are representative of the prison distribution in the world. They are shown in the table below:

Table 1.3: French prison population in 2018, as a percentage of total

Intentional homicide	**Violent or sexual crimes**	**Corruption and Economic Crimes**	**Human Trafficking**	**Firearm Trafficking**	**Drug Consumption and Trafficking**
9.97%	31.32%	29.32%	4.14%	5.63%	19.62%

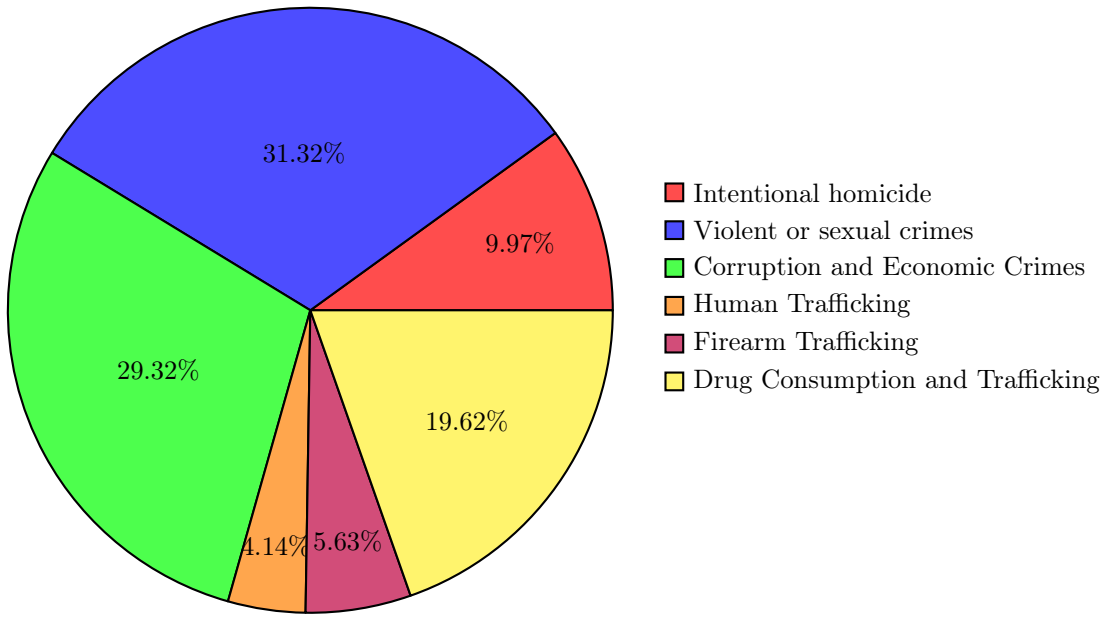

Figure 1.7: French prison population in 2018, as a percentage of total

Using these figures to determine the weights associated with each variable, we obtain the following map, representing the crime index by country:

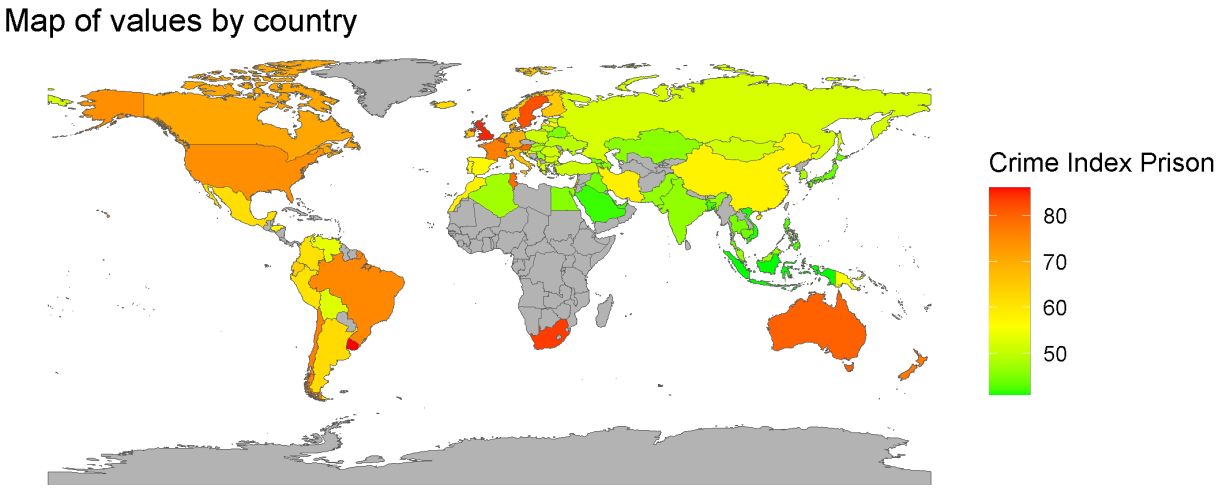

Figure 1.8: Map of the crime index by country, using the prison population method

This map has the merit of offering a different distribution of the weights of the different crime categories, which makes for a slightly different representation. Indeed, as can be seen in Figure 1.3, this index gives much more importance to economic crimes and violent and/or sexual crimes. This will allow us to have different interpretations with our econometric model.

In order to enrich them, we have also decided to take the index created by Numbeo, mentioned earlier, despite these flaws.

1.3.4 Insecurity-based index

The Numbeo Crime Index is an index that measures the level of crime in cities and countries around the world. It is based on data reported by Numbeo users.

The index is calculated using a formula that takes into account reported crimes such as robbery, assault, burglary, car theft, vandalism and terrorist acts. The formula also uses the size of the population to calculate the crime rate per 100,000 inhabitants.

The index is expressed on a scale of 0 to 100, where 0 represents no crime at all and 100 represents a very high level of crime. Based on the score, the following map is obtained:

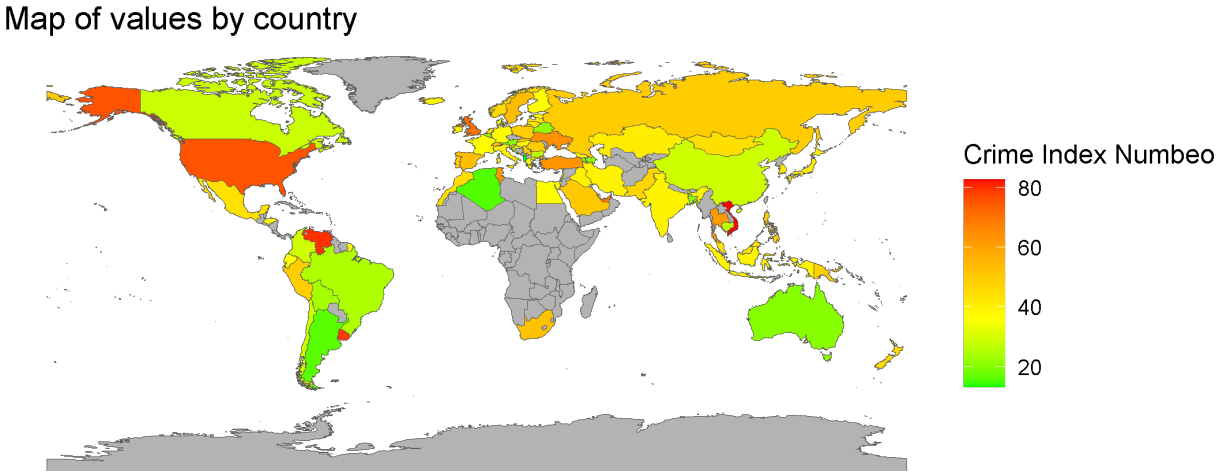

Figure 1.9: Map of the crime index by country, using the Numbeo method

We can see from this map 1.3.4 that the quality of the index is questionable, but we may be able to conclude something interesting about insecurity and the factors that influence it.

1.4 Limits of our work

In this section, we will only discuss the limitations related to our construction of the crime indexes and the collection of the associated variables. We will return later to the limitations of our research work in more general terms.

First of all, one of the first problems we have is the fluctuating definition of crime. We have seen that there are a thousand and one ways to characterise crime in a country.

To try to solve this problem, we chose 82 countries because of their geopolitical proximity. Indeed, these countries share a number of common points, in particular that of being a sovereign state, which is capable of fighting crime, and which has a definition of crime similar to that of the UN.

Despite this selection, the countries remain very different from a legislative point of view, but we needed enough countries to create an interesting sample to analyse.

Still in relation to countries, some may have lied about some of the figures given to the UN, in order to appear more in control of delinquency in their country. This could diminish the quality of our data, but we cannot know to what extent.

For the variables selected, we took the most important and complete ones. Most OECD countries had provided a lot of data to the UN, but for some smaller and less developed countries, some categories were severely lacking in detail and data. We therefore had to restrict ourselves to these 13 criminal variables to create our indices.

Concerning the index created from the prison population, we considered that France's prison population in 2018 was representative for characterising the global prison distribution. This may slightly diminish the quality of this index and thus its interpretations.

Finally, in this study, we have taken the definition of crime in 2018 and it would be very irrelevant to apply this to the past or the distant future. The method would probably still be good, but all the figures would have to be updated, especially those used to calculate the index weights.

Finally, in this chapter we have seen the process of creating three crime indices:

- one based on the total cost of crime to society
- one based on prison distribution
- one based on a mixture of survey and data from Numbeo

First, we proposed several definitions of crime, before choosing one that is relevant to the purpose of this study

Second, we discussed how we collected the data needed to create these indexes. We also discussed the choice of the 82 countries and their consequences.

Finally, we recalled the limitations of our work, due in particular to the choices we had to make in our research.

We are now going to start the data analysis part, starting with a part essentially based on Data Science. For this we will use the statistical software R to observe correlations between our data and try to predict some results of our future regressions.

Chapter 2

Data Analysis

In this chapter, we will apply the different methods of data analysis in order to extract as much information as possible from our collected data.
We will use Principal Component Analysis, or PCA, and clustering, particularly with the use of the k-means method.

We will focus on the analysis of crime and socio-economic data to understand the factors that contribute to crime rates. Using PCA, we can identify the key variables that explain the variation in crime rates in different regions. We can then use k-means clustering to group regions according to their crime rates and other relevant factors. The analysis of socio-economic data also allows us to identify socio-economic factors that may influence crime rates.

Finally we will take all our data and indices and try to use the results to predict the outcome of our regressions.

Ultimately, we will combine all our data and indices to develop a regression model that can predict crime rates based on a range of socio-economic factors. This will provide valuable insights for law enforcement agencies and policymakers, allowing them to develop targeted strategies to reduce crime rates and improve community safety.

In summary, this chapter will provide a comprehensive overview of data analysis techniques and their application in the field of crime and socio-economic analysis. By combining these techniques, we can uncover hidden patterns and relationships in our data, leading to valuable insights and predictions that can inform policy decisions and improve public safety.

2.1 Criminal Data

Crime analysis is a major concern for many cities and countries around the world. Governments and researchers seek to understand crime trends in order to take measures to prevent and reduce them. This study focuses on the correlations between different categories of crime. In this section, we will address the crime data. There are 13 crime data sets:

- Use and traffic of various drugs: Cannabis, Cocaine, Amphetamines (including amphetamines and methamphetamines), Ecstasy and its derivatives, Opioids (including prescribed opioids and contraband), Opiates, Tranquilizers and Sedatives (including barbiturates and benzodiazepines) and New Psychoactive Substances or NPS (mainly including ketamine)
- Intentional homicide
- Violent and/or sexual crimes (including acts intended to induce fear or emotional stress, child pornography, kidnapping, robbery, violence against the person, sexual exploitation and all categories of sexual violence)

- Economic crimes (including illegal waste dumping, environmental pollution, burglary, corruption, fraud, money laundering, migrant smuggling and all categories of theft)
- Firearm trafficking (including the number of weapons seized by justice, customs authorities and returned to the police)
- Human trafficking

2.1.1 Correlation Analysis

The aim of this correlation analysis is to determine whether there are significant relationships between these different crime categories, in order to better understand crime trends and to be able to take measures to combat them more effectively.

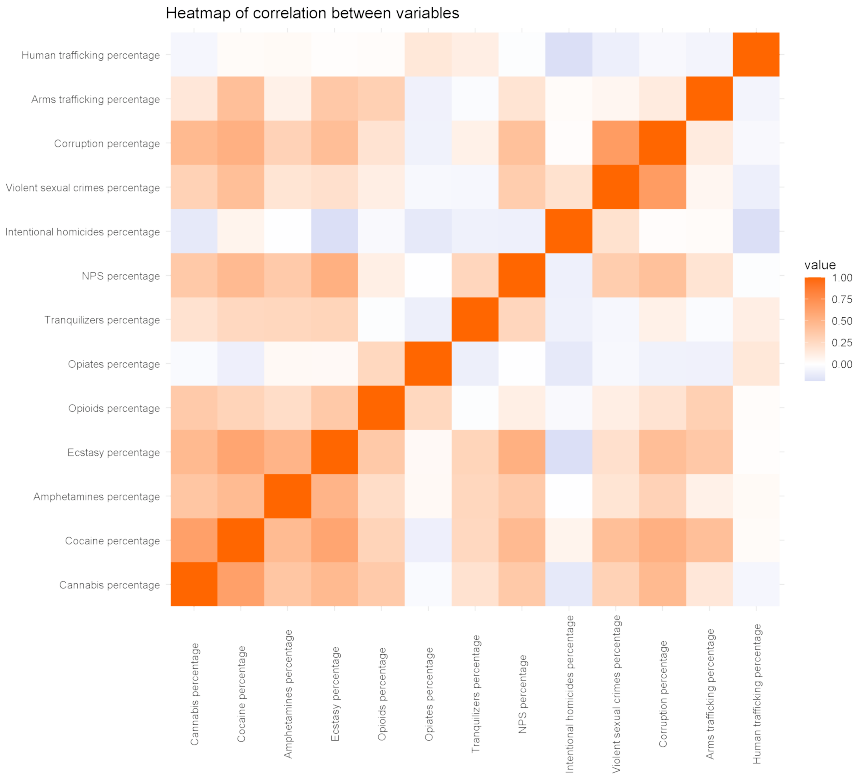

Figure 2.1:

It can already be seen that the different drugs are strongly correlated with each other, except for opiates, which are only weakly correlated with all other variables.

This correlation suggests that people who use one drug are more likely to use another drug. This may be due to common factors such as age, social environment, availability of the drug, personality characteristics, mental health disorders, etc. However, for opiates, the correlation is not as strong as for other drugs.

However, with regard to opiates, the correlation is less strong. Opiates, such as morphine, oxycodone, heroin, etc., are drugs that act on the central nervous system to relieve pain and produce effects of well-being and euphoria. These drugs are often prescribed to treat severe pain, but can also be used illicitly.

Opiates have a strong capacity to cause physical and psychological dependence, which means that users can quickly become addicted to these drugs. This can lead to regular use of opiates in high doses, which can make users less interested in using other drugs.

Secondly, it can be noted on the figure 2.1.1 that the statistics on opiate consumption, the number of homicides and human trafficking are very much 'on their own'. Indeed, they seem to have very little correlation with other crime statistics.

This can be explained by the fact that most homicides are not related to drug use, but rather to personal conflicts, crimes of passion, criminal activities, etc. Therefore, although drug use may be a factor contributing to the increase in violence, it is only a small part of the overall factors influencing the number of homicides.

Furthermore, drug use may affect behaviour and impulses, but it cannot alone explain the complex motives behind homicides.

On the other hand, opiate and human trafficking can be considered as 'niche' trafficking. Consequently, they are far removed from the 'classic' trades, which may explain this correlational isolation.

There is also a very strong similarity between economic crime and violent or sexual crime. These are very strongly correlated with each other and are linked to drug use.

Indeed, certain economic crimes such as financial fraud or money laundering can be linked to drug use, which can lead individuals to seek quick and illegal gains to finance their consumption. Similarly, drug use can sometimes make individuals more impulsive or aggressive, thus increasing the risk of committing violent or sexual crimes.

However, the correlation between drug use and economic, violent or sexual crime does not necessarily mean a direct causal relationship, but rather a statistical association.

In order to analyse these relationships in more detail, we will now use a Principal Component Analysis, or PCA, to take the investigation further.

PCA is a multivariate statistical method used to reduce the dimensionality of data and explore relationships between variables. It consists of transforming a set of correlated variables into a set of uncorrelated variables called "principal components". These components are ordered according to their respective contribution to the total variance of the data.

The objective of PCA is to find a representation of the data in a new lower-dimensional space that captures most of the information contained in the original variables, while reducing noise and the effect of redundancy between variables. This representation allows the relationships between variables to be visualised graphically, the most important variables to be identified and groups of similar variables to be detected.

PCA is often used for data pre-processing prior to the application of other statistical modelling or machine learning techniques, or for the visualisation of complex data.This is the reason why we will apply it now.

First, let's look at the eigenvalues of our PCA. Eigenvalues are an important concept in principal component analysis. When a PCA is performed, the data are transformed into principal components that are linear combinations of the original variables. Eigenvalues are a measure of the variance explained by each principal component. They represent the amount of total variance in the data that is explained by each principal component.

The elbow method is a commonly used method for determining the number of principal components to be retained in a PCA. It involves plotting the eigenvalues in descending order and observing where a significant change in slope occurs. The number of principal components corresponding to this location is often chosen as the number of principal components to retain.

Specifically, the elbow method involves plotting the number of principal components on the x-axis and the corresponding eigenvalues on the y-axis. Then the inflection point is sought where the slope of the curve decreases sharply, resembling a elbow. This point is considered as the optimal number of principal components to retain. We have plotted the eigenvalues obtained in the table below:

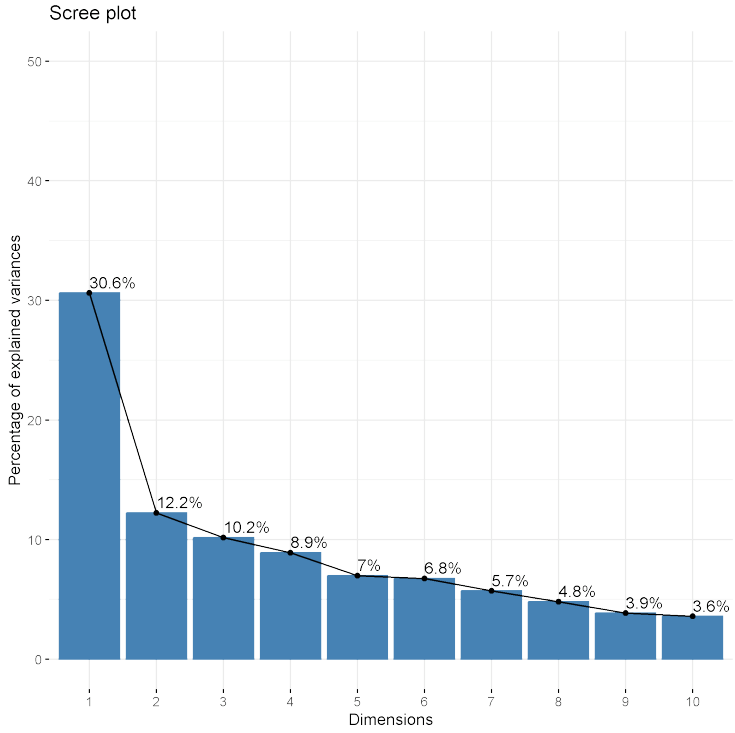

Figure 2.2:

In this table 2.1.1 can be seen that the first two axes explain about 43% of the total variance, while the third axis explains only 10% of the variance. Looking at the curve, we can see that the slope of the curve changes significantly after the second axis, which indicates that the addition of further axes does not contribute significantly to explaining the remaining variance.

Therefore, the choice of keeping only the first two axes is justified by the elbow method, as the first two axes explain a significant proportion of the variance, and adding additional axes would not provide significant additional information.

Now that we have chosen the number of principal components to keep, let's analyse the results. Here is the circle of correlations of the variables:

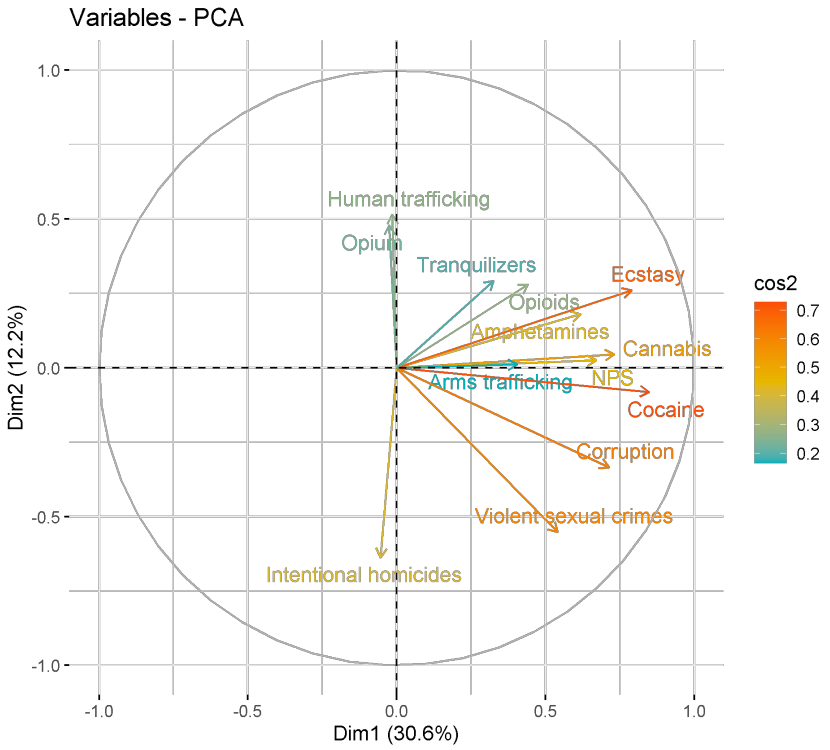

Figure 2.3:

On this graph, we have also plotted the cos2 of each variable. The cos2 (square of the cosines) is a measure of the quality of representation of the variables on the factorial axes in a principal component analysis. The cos2 is used to evaluate the share of variance of the variable explained by each factorial axis.

More precisely, the cos2 represents the proportion of the total variance of the variable that can be explained by the axis under consideration. It is the square of the cosine of the angle formed between the variable and the factorial axis. Thus, a cos2 close to 1 indicates that the variable is well represented on the corresponding factorial axis, while a cos2 close to 0 indicates a weak representation of the variable on the axis.

The cos2 can therefore help to interpret the PCA results by allowing the identification of variables that have a strong influence on each axis. Variables with high cos2 for a particular axis are those that have a significant contribution to that axis and are therefore the most relevant for the interpretation of the axis.

First of all, it can be seen that on this graph, the values with the lowest cos2 are the variables isolated from the others: Human and Weapons Trafficking and Consumption of Opiates, Tranquillizers and Opioids. In other words, they do not contribute much to understanding the relationships between the other variables.

On the other hand, all the other variables have a very high cos2 , which means that all the variables have a good quality of representation on the first two factorial axes, which justifies the choice to keep only these two axes. This is confirmed by the fact that the sum of the cos2 of the variables on these two axes represents more than 70% of the total variance of the data.

Now let's analyse the other information that this graph gives us.
First of all, we can note that the information given by the correlation matrix is confirmed by this circle. Indeed, we can see that the variables of human trafficking and opiates are isolated but together and that voluntary homicides are completely separate.

Thanks to this graph 2.1.1, we can go further: we can affirm that these three statistics are independent of the other criminal variables.

Secondly, we can observe a convergence of the statistics representing drug consumption. We can therefore conclude that the use of the following drugs is closely related: Cocaine, Cannabis, NPS, Ecstasy and Amphetamines.

We can also see that economic crimes and violent and sexual crimes are very close. These two statistics are therefore strongly correlated.

Now that we have analysed the statistical relationships between our crime variables, let us turn to the analysis of the individuals in our sample.

In the graph below, we have plotted the representational qualities of each individual on our two axes:

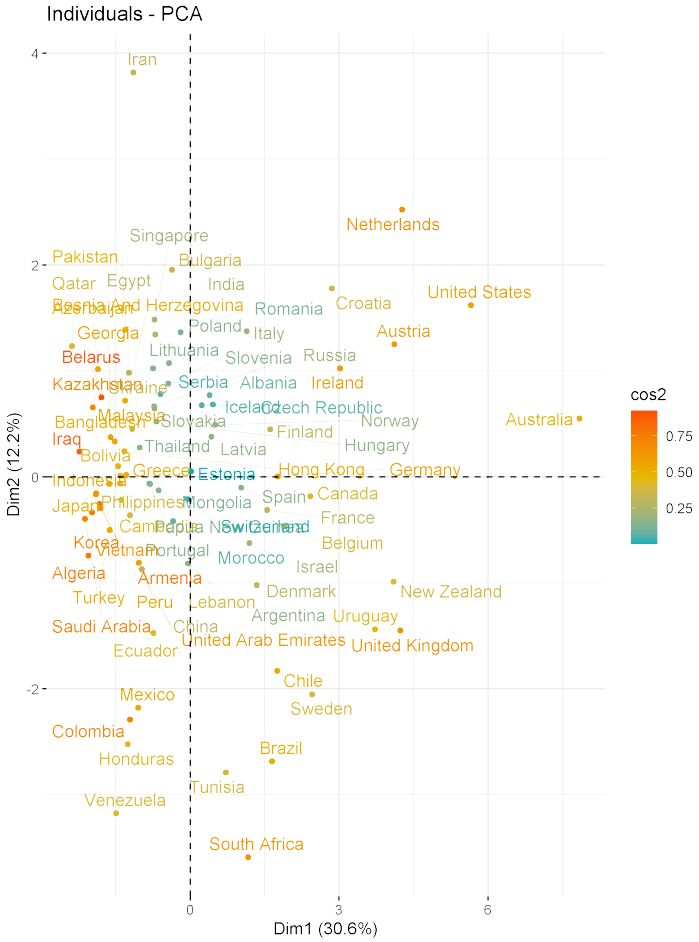

Figure 2.4:

This representation shows that some countries stand out, notably Australia, the United States, the Netherlands and Austria in terms of drug use. There are also unusually high figures for countries such as South Africa, Venezuela, Brazil, Tunisia and Honduras.

Unfortunately, the graph is not very readable due to the large number of countries analysed.

We will therefore use clustering methods to better visualise the relationships between countries and crime statistics.

2.1.2 Clustering

Clustering methods are unsupervised data analysis techniques that group similar observations into classes or clusters. Clusters are subsets of observations that have common characteristics and are different from other clusters.

There are different clustering methods, each with its advantages and disadvantages:

- Hierarchical clustering: this method consists of building a hierarchical cluster structure by merging the most similar clusters, until all points are grouped into one cluster. The advantages of this method are that it is easy to interpret, it allows the cluster structure to be visualised and it does not require the number of clusters to be specified in advance. The disadvantages are that it is sensitive to outliers and can be very time consuming for large databases.
- K-means: this method consists of specifying the number of clusters in advance and finding the centres of these clusters iteratively. The advantages of this method are that it is fast, suitable for large datasets and easy to implement. The disadvantages are that it can be sensitive to randomly chosen starting centres and that it requires the number of clusters to be specified in advance.
- Density clustering: This method involves finding dense regions of points in the data space, separated by less dense regions. The advantages of this method are that it is robust to outliers, can identify non-spherical shaped clusters and does not require the number of clusters to be specified in advance. The disadvantages are that it can be sensitive to the choice of density parameters and can be computationally expensive for large databases.

Clustering methods are useful for exploring data, detecting hidden structures and identifying groups of similar observations. They are used in many fields, such as biology, finance, geology, image and signal processing, psychology, marketing and many others. Applications of clustering are numerous, for example in market analysis to segment consumers into homogeneous groups, in medicine to identify subgroups of patients with similar disease profiles, or in forensics to detect patterns of crime.

Unfortunately, with our data, we cannot use the hierarchical clustering method since we have too much extreme data which prevents the effective use of this technique.The density clustering method also proved ineffective.

We will now apply the k-means method. To do this, we need to establish the number of clusters we want.

To this end, we will again apply the elbow method, but this time apply it to the intra-cluster distance.

The graph below represents this distance, as a function of the number of clusters:

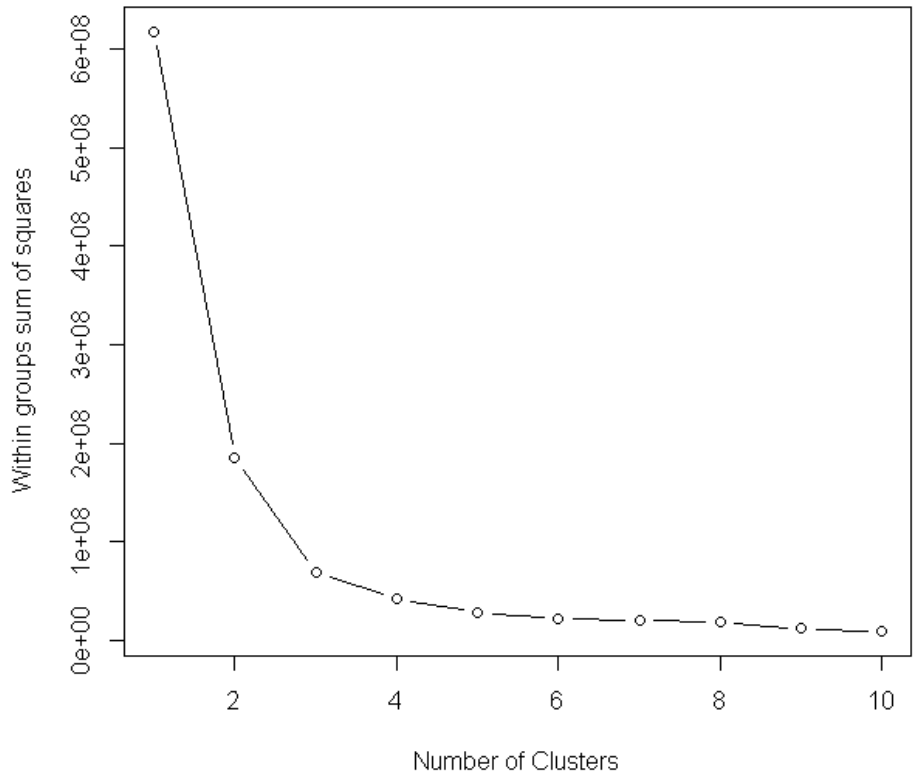

Figure 2.5:

Using the elbow method, we can conclude that this sample of countries can be grouped into 3 clusters. We can determine these with the k-means method.

This gives us the following representation:

Figure 2.6:

The graph shows the formation of three groups:

- The first is made up of Australia only
- The second is made up of the following countries: USA, Austria, Italy, Finland, Norway, Switzerland, Germany, France, Canada, Belgium, Denmark, Chile, Sweden, Tunisia, South Africa, UK, Uruguay and New Zealand.
- The third includes all remaining countries

Let us now see how these groups are characterised, using our criminal variables:

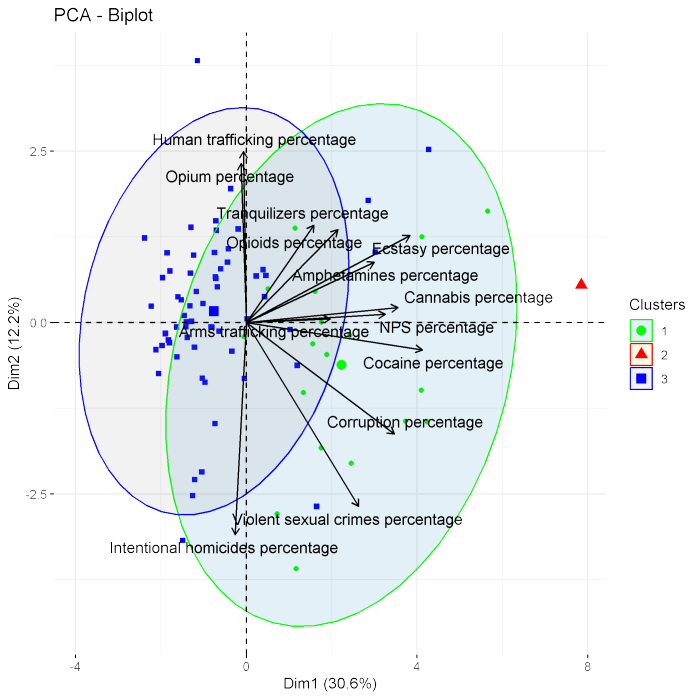

Figure 2.7:

From this graph 2.1.2 we can conclude several things:

- Australia is a very special country, because of its uncommon rate of drug use of all kinds.
- The second group, rather composed of rich and developed countries, is characterised by a rather high crime rate. This is very much centred on the use of various drugs, but not in the same way as in Australia.
- The third group, composed of the rest of the less developed countries, is quite disparate. Here we find countries with very low crime and those with 'special' crime. Here we find countries with a high rate of homicide, trafficking in human beings and opium.

By studying crime statistics, we have seen that crime is not a homogeneous phenomenon, but is divided into different categories. We have seen that some countries have very high crime rates related to drug use, while other countries are characterised by more violent crime, such as homicide and human trafficking.

These differences are important to understand because they allow for more appropriate crime control strategies to be developed.

We have been able to express and categorise the different types of crime and illustrate these different categories with representative groups of countries.

We have also been able to highlight the statistical relationships between the different crime variables and their implications.

We will now turn to the analysis of the socio-economic characteristics of the countries, in order to better understand the phenomena and correlations in this area.

2.2 Socio-Economic Data

Macroeconomic data are at the heart of our subject. Indeed, we believe that they have an impact on the whole society, including crime. Before trying to model this relationship, it is important to know more about these data and their characteristics.

First, we will justify our choice of variables.

2.2.1 Which variables we chose and why ?

For this study we selected 24 socio-economic data:

- "Net migration": This variable represents the net migration of a country as a percentage of its total population, i.e. the difference between the inflow and outflow of migrants. It is important because it can have an impact on cultural diversity and social tensions in a country, which in turn can affect the crime rate.
- "Religion": The variable Religion refers to the religious composition of a country. Within this, we counted the number of religions practised in the country with at least 1% of the total population. This variable can have an influence on inter-community relations, as well as on the values and norms of society, which can also affect the crime rate. It can also generate strong tensions between different communities.
- "Justice": This variable represents the efficiency of a country's justice system, measured by the percentage of GDP allocated to the Ministry of Justice. This variable is important because it can indicate a country's ability to prevent and punish crime. It also reflects a willingness to follow up with a view to the eventual reintegration of the criminal into society.
- "Police": The Police variable refers to the percentage of GDP allocated to the Ministry of the Interior. It is important because it can have an impact on the ability of the police to prevent and solve crimes. It may also characterise police states that are more concerned with punishing their population than their security.
- "Education: This variable represents the percentage of GDP allocated to the Ministry of Education. Education is important because it can have an impact on economic opportunities, income levels and social attitudes, which in turn can influence crime rates. It can also be assumed that a well-performing school produces citizens, who are expected to respect the law, thus reducing the number of criminals.
- "Prison": The Prison variable represents the percentage of GDP allocated to the management of prisons. It is important because it can indicate the severity of a country's judicial system and its ability to prevent crime. More importantly, it provides an indicator of how criminals in a country are treated.
- "Death Penalty: This variable represents whether or not a country has the death penalty. It is important because it can have an impact on attitudes towards crime and the deterrence of crime.
- "Unemployment rate: This variable represents the percentage of the labour force that is unemployed. Unemployment can have an impact on poverty and economic opportunities, which in turn can influence the crime rate.
- "Poverty": The Poverty variable represents the percentage of the population living below the poverty line. Poverty can have an impact on economic opportunity and social despair, which in turn can influence the crime rate.
- "Homelessness": This variable represents the percentage of homeless people in a country. It is important because it can indicate the levels of poverty and social problems in a country, which in turn can affect the crime rate.

- "Malnutrition" : This variable represents the percentage of the population that is malnourished. It is important because malnutrition can have an impact on physical and mental health, which in turn can affect crime rates.
- "Mental health": This variable represents the number of mental health care providers per 100,000 population. Mental health can have an impact on behaviours and attitudes, which in turn can affect the crime rate.
- "Drugs": This variable represents the legislation against drug use ranging from 0 to 5, from the death penalty to the legalisation of most drugs. This undoubtedly influences drug use and therefore crime.
- "Democratic" : This variable measures the level of democracy in the country, on a scale of 1 to 10. Studies have shown that countries with higher levels of democracy tend to have lower crime rates.
- "Homeowner": This variable measures the percentage of the population that owns their home. Studies have shown that homeowners tend to be more involved in their community and therefore contribute to lower crime.
- "Students': This variable measures the total enrolment in higher education regardless of age, expressed as a percentage of the total population in the five-year age group following the completion of secondary education. Studies have shown that cities with higher student rates tend to have lower crime rates.
- "Illiterate": This variable measures the percentage of the population that is illiterate. Illiterate people may be more vulnerable to crime and may be involved in criminal activities.
- "Gini index": This index measures income inequality in a country. Studies have shown that countries with higher levels of income inequality tend to have higher crime rates.
- "GDP": This variable measures the gross domestic product (GDP) of the country. Studies have shown that there is a negative correlation between the level of GDP and the crime rate.
- "Inflation": This variable measures the inflation rate in the country. Studies have shown that higher inflation rates can be associated with higher levels of crime.
- "Median age": This variable measures the median age of the population. Studies have shown that younger populations tend to have higher crime rates.
- "Urbanisation rate": This variable measures the percentage of the population living in urban areas. Studies have shown that cities tend to have higher crime rates than rural areas.
- "Population": This variable measures the size of the country's population. The larger the population, the more likely it is to have high crime.
- "Immigrants": This variable measures the percentage of the population that is foreign born. Immigrant populations may be more vulnerable to crime and may be involved in criminal activities.

The variables selected for our study all have a potential link with crime. For example, the unemployment rate may cause some people to commit crimes to support themselves. Similarly, the quality of the justice and prison system may influence the crime rate by discouraging potential offenders or preventing them from re-offending. The level of poverty and malnutrition can also have an impact on crime by creating socio-economic conditions conducive to offending.

Religion can also play an important role in crime, particularly in relation to inter-religious conflict. Drug policies and mental health are also important variables, as drug use can lead to criminal behaviour and mental health problems can influence the actions of individuals.

Other variables such as median age, Gini index, GDP, inflation, the proportion of immigrants and the rate of urbanisation can also have an impact on crime. For example, a high proportion of immigrants can lead to social and ethnic tensions that can lead to criminal behaviour, while economic inequality and poverty can lead to violence.

In sum, each variable selected for our study has a potential link to crime and can help us to better understand the factors that contribute to its increase or decrease.

2.2.2 Correlation Analysis

The aim of this correlation analysis is to determine whether there are significant relationships between these different statistics, in order to better understand the relationships between them and thus the phenomena underlying the economy and society.

With the data collected, we obtain the following correlation matrix:

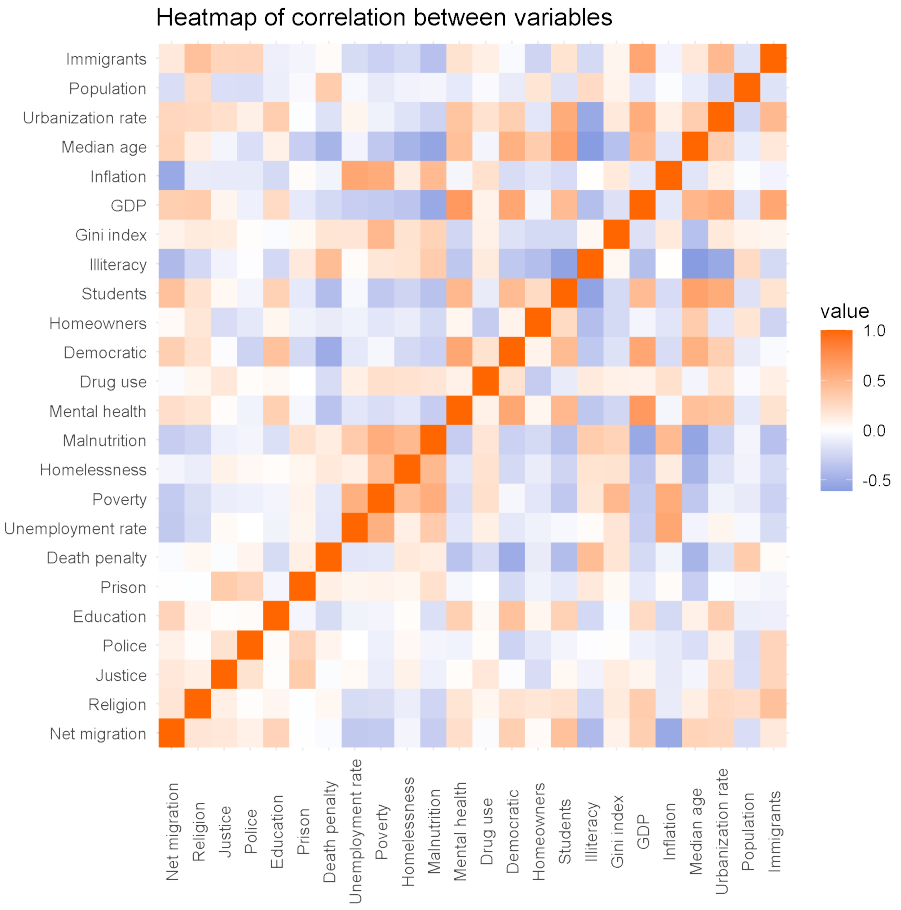

Figure 2.8:

In view of the large number of different categories, we will keep the analysis of this matrix fairly concise. We will come back to the relationship with the correlation circle in more detail.

First of all, let us note the variables that are very poorly correlated with the others. We see that the government budget statistics (Justice, Police, Education and Prison) are not correlated at all with the other economic data. This can make us question the efficiency and relevance of each ministry. Indeed, what is the point of increasing the budget of these if the effects on the country are almost non-existent?

This may also suggest that policies to reform the institutions of justice, police and the prison system may be needed to make them more effective and fair in their operation.

There is also a correlative isolation of religion from other statistics. This may reflect a distancing of religious institutions from political and social decision-making in developed countries. In other words, it may reflect a trend towards secularisation in developed countries, where religion is losing its influence in the political and social spheres. It may also show that religion is flourishing, regardless of socio-economic conditions.

Similarly, the death penalty statistic has very little correlation with most macroeconomic statistics, except for the democracy index. This makes sense: whether or not a country has the death penalty has little impact on its economy, but it does indicate a certain tendency towards authoritarianism. Indeed, countries that have abolished the death penalty often have a more democratic political system that respects human rights. Conversely, countries that continue to apply the death penalty can be considered authoritarian or conservative.

Another group of similar variables can be observed: unemployment, poverty, homelessness and malnutrition. As one would expect, these 4 statistics are highly correlated with each other. We can also see that they are negatively correlated with GDP and median age. On the one hand, it is possible that poverty and economic insecurity lead to an increase in the number of homeless or malnourished people. On the other hand, unemployment can have a direct impact on poverty by reducing people's incomes.

However, it should be noted that the relationship between GDP and these variables cannot be considered causal. Indeed, although GDP is an important indicator of a country's economic health, it does not take into account factors such as the distribution of wealth or the quality of life of citizens. Thus, it is possible that countries with high GDP still have high rates of poverty or malnutrition due to specific structural or political factors.

We will now analyse these relationships in more detail, using a PCA. We must therefore choose the number of dimensions we will keep. For this we will again use the kink method, based on the eigenvalue graph, shown below:

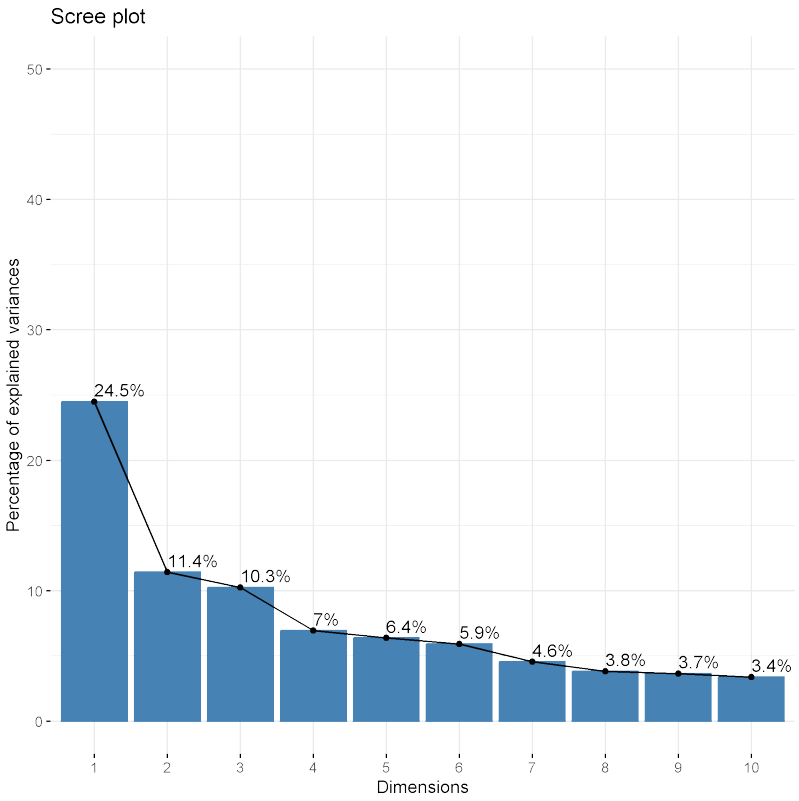

Figure 2.9:

In this table 2.2.2 can be seen that the first two axes explain about 36% of the total variance, while the third axis explains only 10% of the variance. Looking at the curve, we can see that the slope of the curve changes significantly after the second axis, which indicates that the addition of further axes does not contribute significantly to explaining the remaining variance.

Therefore, the choice of keeping only the first two axes is justified by the elbow method, as the first two axes explain a significant proportion of the variance, and adding additional axes would not provide significant additional information.

Now that we have chosen the number of principal components to keep, let's analyse the results. Here is the circle of correlations of the variables:

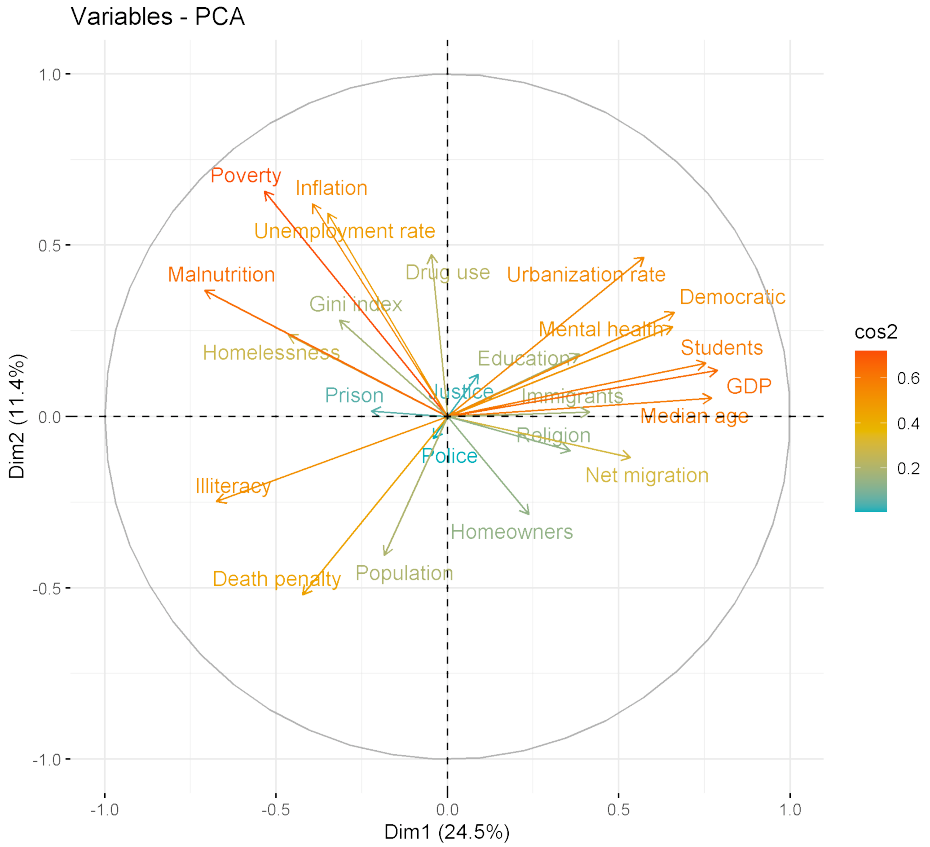

Figure 2.10:

First, we note the variables with a low cos2 , i.e. poorly represented by the PCA. In this category we find statistics on government budgets, population, religion, immigration, home ownership, the Gini index and drug legislation.

These statistics have little influence on the total variance of the data. It is important to note that this does not necessarily mean that these variables are less important than the others. Their low correlation may simply indicate that they are less strongly related to other economic variables in the data analysed. For example, religion may play an important role in people's lives, but this may not be directly reflected in the economic data we have analysed. Similarly, drug legislation may have important effects on crime, but may not be directly related to the other economic variables we have studied.

For the sake of clarity we will redo the correlation circle without the variables mentioned above, which are not needed in this analysis.

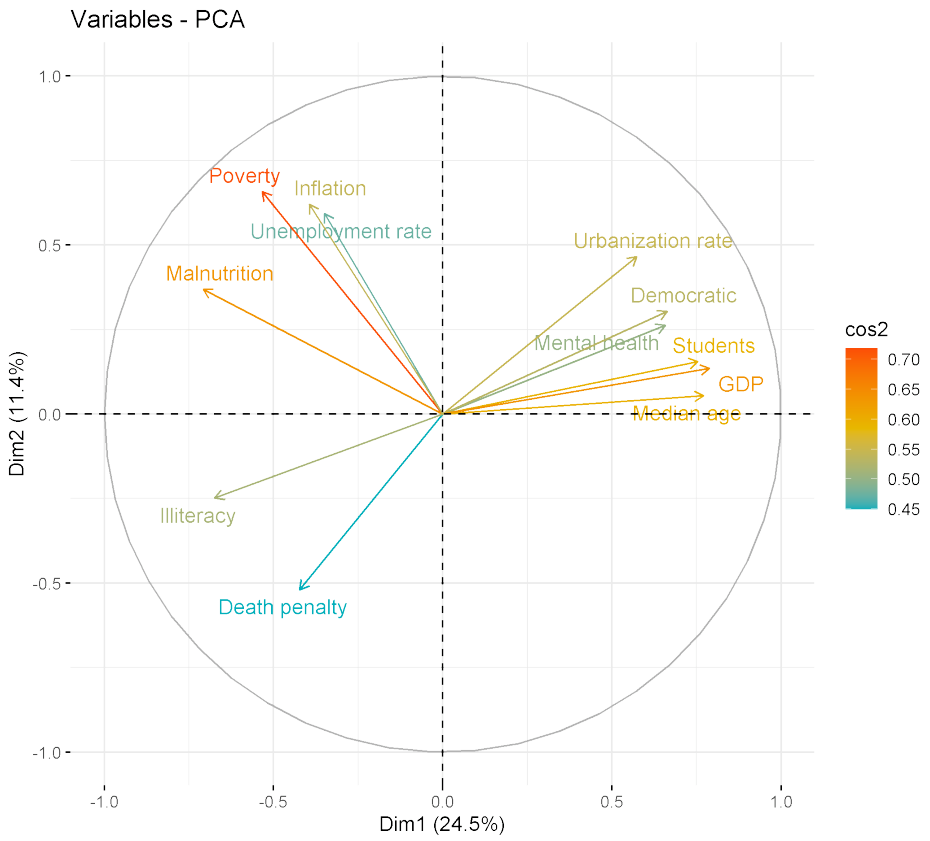

Figure 2.11:

Thanks to this new graph, we can better see the relationships between the main variables of our model.

We can already confirm the isolation of the death penalty. It is unlikely to be related to the socio-economic statistics of a country.

Two very homogeneous and compact groups can be observed:

- The first is all the "negative" statistics: Poverty, Inflation, Unemployment and Malnutrition. All these variables are highly correlated with each other, but very little with the rest of the statistics.
- The second is composed of all the "positive" variables: Urbanisation rate, Democratic Index, Mental Health, students in higher education, GDP and median age. All these variables are highly correlated with each other and completely independent of the first group.

It can be noted that the literacy rate is strongly negatively correlated with this last group.

The group of negative variables, which are strongly correlated with each other, can be considered as an overall indicator of poverty and insecurity in a country. It is therefore important to take this indicator into account when analysing crime, as poverty and deprivation may be factors contributing to the commission of criminal acts.

On the other hand, the group of positive variables can be seen as an overall indicator of development and quality of life in a country. It is therefore important to also take these variables into account when analysing crime, as a more developed country with a better quality of life can potentially have lower crime.

Finally, the strong negative correlation between the literacy rate and the group of positive variables can be explained by the fact that countries with lower literacy tend to be less developed and offer a lower quality of life, which is reflected in the variables of the second group. This also highlights the importance of education in the development and quality of life of a country, as well as in the prevention of crime.

Now that we have analysed the statistical relationships between our crime variables, let us turn to the analysis of the individuals in our sample.

In the graph below, we have plotted the representational qualities of each individual on our two axes:

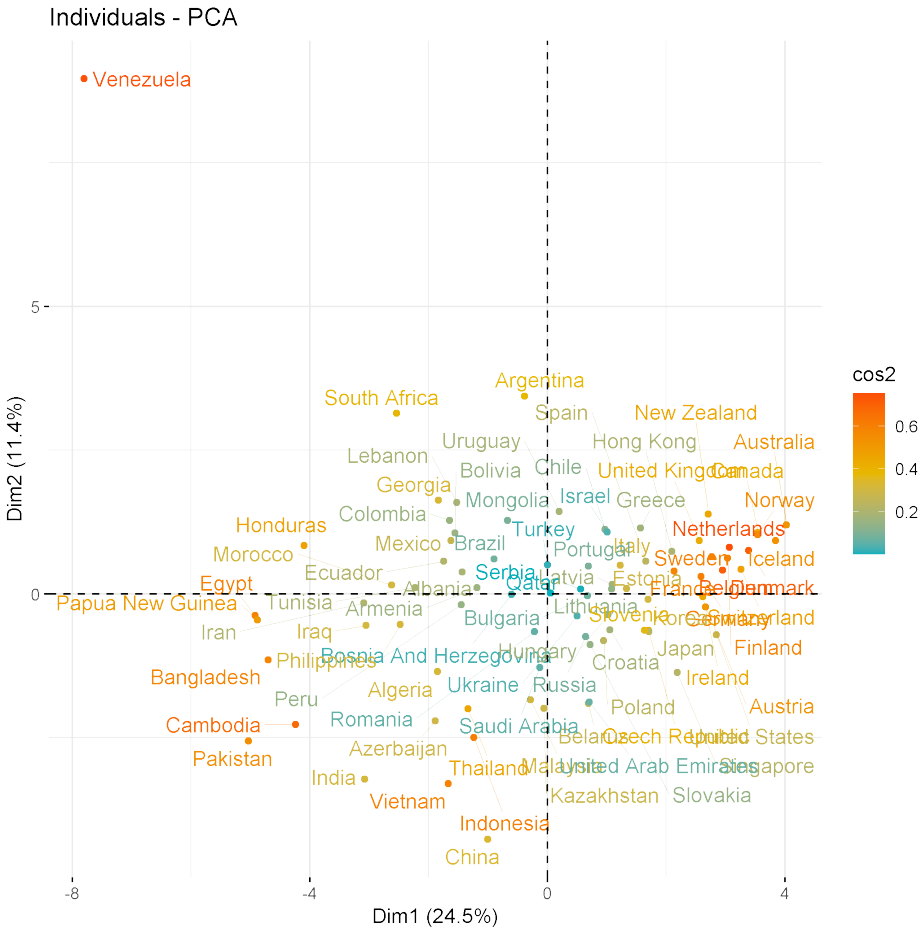

Figure 2.12:

This graph shows the singularity of Venezuela. Indeed, it is very far from the others on the two main axes. This shows that this country is very different from the others at the macroeconomic level.

Apart from this country, the other countries form a fairly compact group, which shows that our work has been well done, we have chosen similar countries from a socio-economic point of view.

There are a few countries that stand out from the crowd, notably Papua New Guinea and South Africa.

Due to the lack of clarity of the graph we cannot do further analysis, so we will use the clustering method to see if we can find groups of similar countries.

2.2.3 Clustering

Again, our data do not allow us to use the hierarchical clustering method. The density clustering method also proved ineffective.

We will now apply the k-means method. To do this, we need to establish the number of clusters we want.

To this end, we will again apply the elbow method, but this time apply it to the intra-cluster distance.

The graph below represents this distance, as a function of the number of clusters:

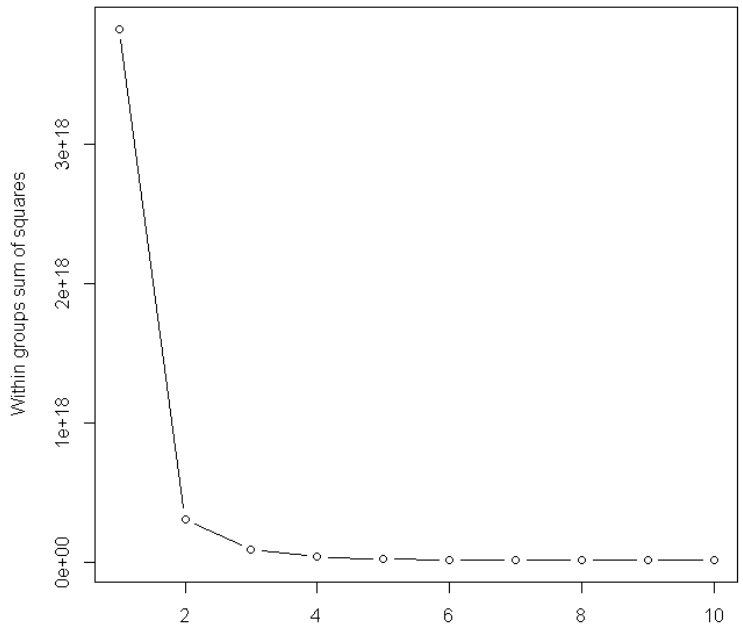

Figure 2.13:

Using the elbow method, we can conclude that this sample of countries can be grouped into 2 clusters. It can be predicted that these two clusters will be composed of only Venezuela for the first and all other countries for the second. If this is the case we will try to create 3 clusters. We can determine these with the k-means method.

This gives us the following representation:

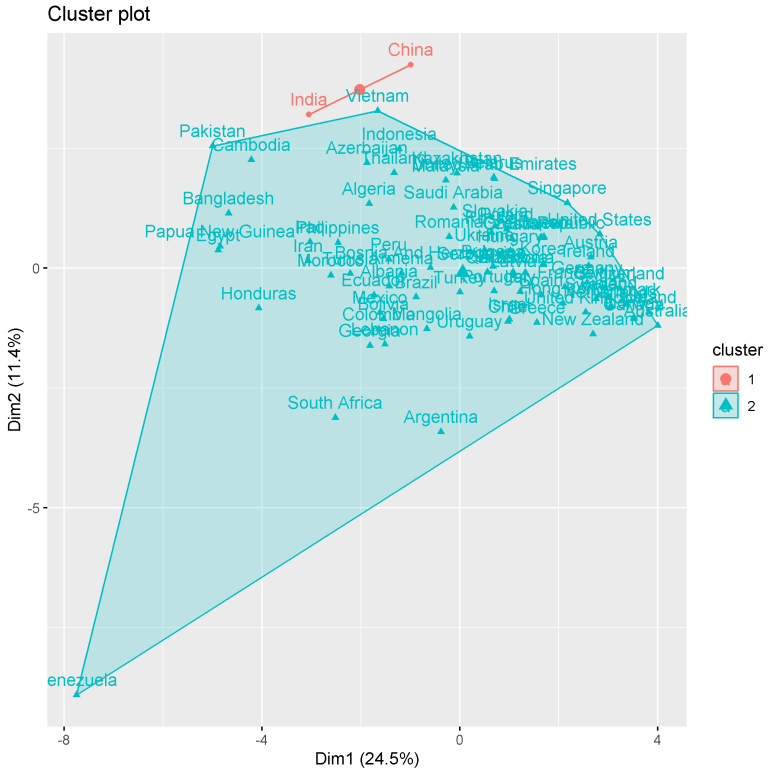

Figure 2.14:

As expected, the graph shows the formation of two groups. However, their composition is somewhat surprising. The first group is composed of China and India. The second group consists of all other countries, including Venezuela.

It is quite logical that China and India are associated, as these two countries are very similar in many macroeconomic statistics, including population.

Let us now see how these groups are characterised, using our socio-economic variables:

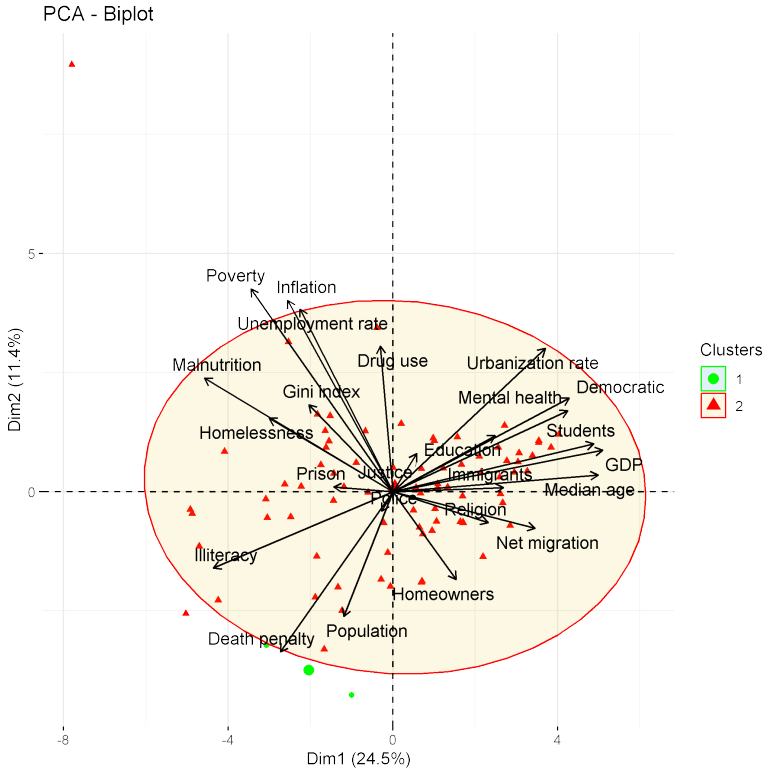

Figure 2.15:

It is clear from this graph what makes China and India a different cluster from the others. These two are distant from the others in 2 variables: death penalty and population. China is the country with the largest population in the world, while India is the second largest. In addition, both countries have extensive use of the death penalty, which differentiates it from other countries. It is therefore these two variables that have contributed to the formation of this particular group.

Apart from these considerations, our sample is quite similar, so we can consider that our choices are relevant for this study. Indeed, the results of our analysis show that the countries we have chosen are similar from a socio-economic point of view. However, it should be noted that the analysis of the data and the relevance of our choices also depend on the assumptions and limits we have set. It is therefore important to take these elements into account in order to interpret the results appropriately.

We will now perform a PCA again to try to find statistical relationships between our criminal and socio-economic variables this time. We may be able to anticipate results that we will get in our econometric model later on.

2.3 Blends of all data types

In this study, our aim is to investigate the factors that influence crime. For this purpose we have mixed socio-economic and crime data in order to observe possible statistical relationships. We will now give the results in the form of a correlation matrix. This can be useful for modelling and for understanding the relationships between the variables. However, it is important to bear in mind that this does not allow us to determine causal relationships between the variables.

That said, the correlation matrix can be a useful tool to identify the variables most strongly related to crime and thus guide the search for causal relationships in econometric analysis.

2.3.1 Correlation Analysis

With the data collected, we obtain the following correlation matrix:

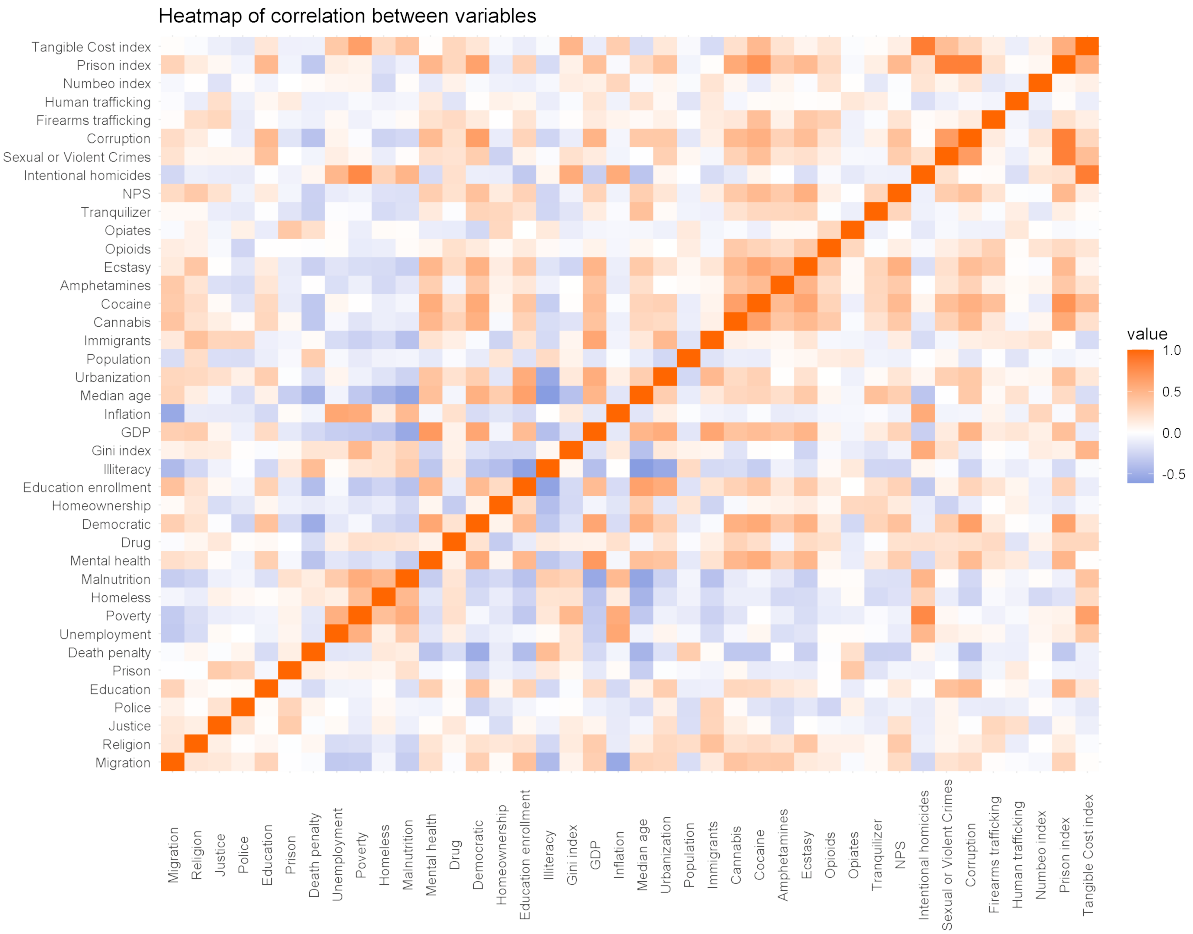

Figure 2.16:

Let's start by analysing the relationships of the indices we have created. First, we can observe that the index proposed by Numbeo is not correlated with any of our statistics. This justifies the basic idea that we had that it was the least interesting of the 3. It is not even correlated with our 2 criminal indices.

This result shows that the Numbeo index is based solely on feelings and not on statistical facts. It therefore measures insecurity only. We may obtain more interesting results with our econometric model.

Indeed, the absence of correlations between the Numbeo index and the other variables may suggest that this index measures a subjective perception of insecurity rather than objective indicators of crime. However, it is important to bear in mind that the absence of correlations does not necessarily mean that there is no causal relationship between the variables.

It will be interesting to see whether the results of our econometric model corroborate this observation and whether the Numbeo index is related to our socio-economic variables.

Let us now turn to the index based on the prison population. It can be seen that the drug with which it is most correlated is cocaine. It is also highly correlated with economic crime and intentional homicide.

It is also highly correlated with our other index, based on tangible cost, which indicates the relevance of the choices of weights in the calculation of our indices.

Indeed, these correlations may indicate that there is a strong association between the prison population and certain types of crime, such as economic crime and intentional homicide. The correlation with cocaine use may also indicate that this drug is often associated with serious crimes that lead to a prison sentence.

In addition, the strong correlation with the tangible cost-based index may indicate that the cost of crime is an important consideration for governments when making decisions about prison sentences. However, as mentioned earlier, it is important to bear in mind that the correlation does not establish a direct causal relationship between these variables.

Now let us turn to our index based on tangible cost. The only remarkable relationship with this index is that it is strongly correlated with the Intentional Homicide.

The fact that the index based on tangible cost shows only one remarkable relationship with intentional homicide can be interpreted in different ways.

On the one hand, it may indicate that the tangible cost of crime (financial cost, infrastructure cost, etc.) is more related to intentional homicide than to other types of crime. This may suggest that policies to reduce the cost of crime should focus more on preventing intentional homicide.

On the other hand, it may also indicate that our tangible cost-based index does not capture all relevant aspects of crime. Indeed, it is possible that some dimensions of crime, such as economic crime, are not captured by this index and therefore do not have a significant relationship with it. This justifies our choice to use several indices in our analysis.

In addition, there is a very strong correlation between poverty and intentional homicide. This suggests that there is a significant relationship between the level of poverty in a country and the rate of intentional homicide committed there.

Firstly, poverty can be associated with situations of stress and frustration, which can increase levels of violence and aggression. Harsh living conditions and economic deprivation can also lead some people to turn to crime to support themselves.

In addition, poverty can also be associated with a lack of access to education, employment and health care, which can contribute to a culture of violence in some deprived neighbourhoods. In these environments, young people can be particularly vulnerable to the influence of gangs and organised crime.

Finally, poverty can also be associated with high levels of insecurity and social tensions, which can foster the emergence of violent conflict. Poor neighbourhoods may also lack the resources to prevent crime and ensure the safety of residents, which can contribute to high levels of crime and violence.

In sum, the correlation between poverty and intentional homicide suggests that there are complex links between socio-economic factors and crime. This underscores the importance of considering these factors when designing public policies to prevent crime and promote safety.

We will now analyse these relationships in more detail, using a PCA. We must therefore choose the number of dimensions we will keep. For this we will again use the kink method, based on the eigenvalue graph, shown below:

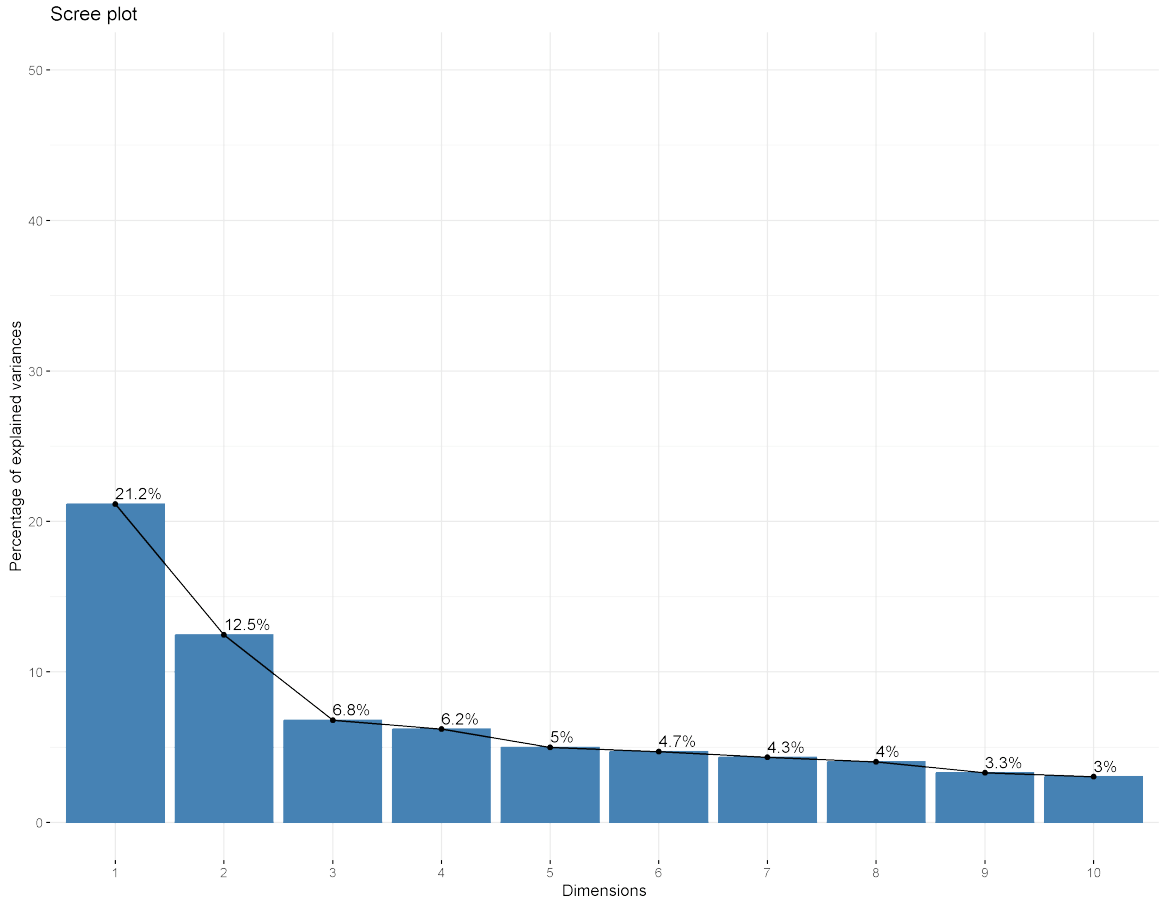

Figure 2.17:

In this table 2.3.1 can be seen that the first two axes explain about 34% of the total variance, while the third axis explains only 7% of the variance. Looking at the curve, we can see that the slope of the curve changes significantly after the second axis, which indicates that the addition of further axes does not contribute significantly to explaining the remaining variance.

Therefore, the choice of keeping only the first two axes is justified by the elbow method, as the first two axes explain a significant proportion of the variance, and adding additional axes would not provide significant additional information.

Now that we have chosen the number of principal components to keep, let's analyse the results. Here is the circle of correlations of the variables:

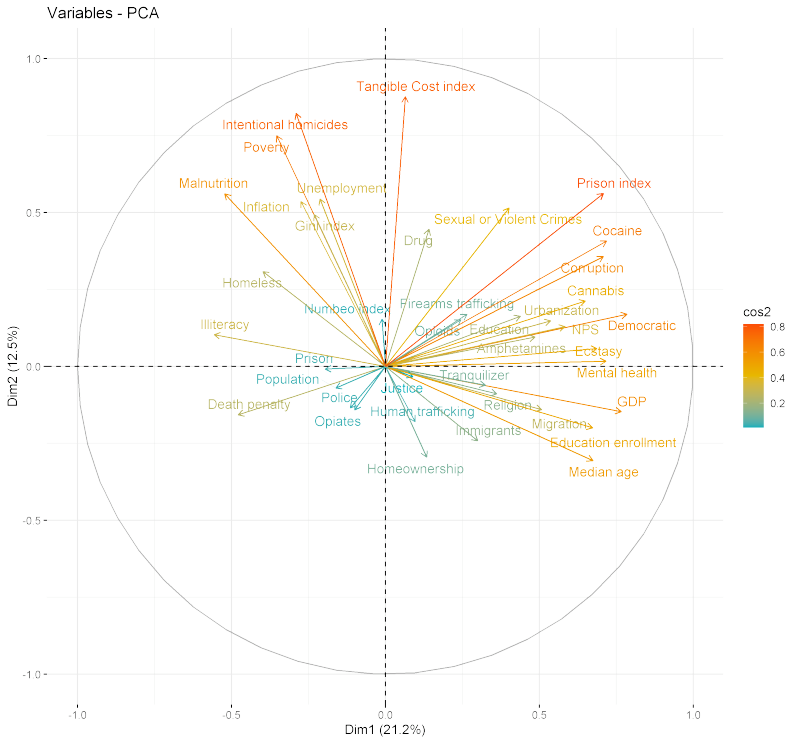

Figure 2.18:

First, we note the variables with a low cos2 , i.e. poorly represented by the PCA. In this category we find statistics on public budgets, population, religion, immigrants, home ownership, opiate and tranquilizer use, arms and human trafficking and the Numbeo index

These statistics have little influence on the total variance of the data. It is important to note that this does not necessarily mean that these variables are less important than the others. Their low correlation may simply indicate that they are less strongly related to other variables in the data analysed.

For the sake of clarity we will redo the correlation circle without the variables mentioned above, which are not needed in this analysis.

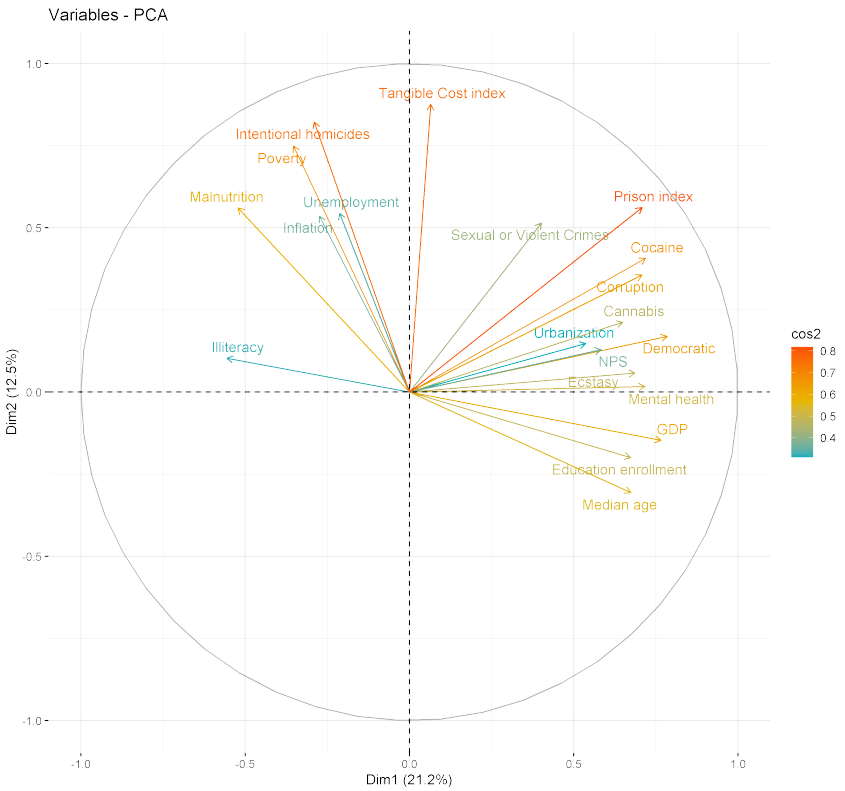

Figure 2.19:

Thanks to this new graph, we can recognise two very homogeneous and compact groups:

- The first is composed of malnutrition, inflation, unemployment, poverty, intentional homicide and the index based on tangible cost.
- The second one is composed of violent and sexual crimes, the index based on the prison population, the use of cocaine, cannabis, ecstasy and NPS, economic crimes, the democratic index, the urbanisation rate, the number of mental health practitioners, the GDP, the number of students in higher education and the median age.

This suggests that socio-economic variables such as poverty, unemployment, inflation and intentional homicide are closely related to each other and can be seen as indicators of the same broader phenomenon.

On the other hand, the crime variables and some socio-economic variables such as GDP, urbanisation and education also seem to be related to each other, but in a different way than the first category of variables.

Finally, it is interesting to note that the index based on the prison population, which was strongly correlated with economic crime and intentional homicide, seems to have a more complex relationship with the other variables. It is strongly related to drug use, particularly cocaine, as well as to socio-economic variables such as GDP and the number of students in higher education. This may suggest that prison overcrowding is related to wider economic and social factors, as well as to crime.

It can be noted that the variable representing illiteracy is opposite to the second group. It is therefore negatively correlated with it. This can be explained by the fact that countries with a high literacy rate generally have a higher level of development and better education, which may positively influence crime and reduce the use of violence as a means of conflict resolution. In contrast, in countries with lower literacy rates, people may have fewer economic opportunities and be more likely to be involved in criminal activities.

The two groups described above can also be seen as independent of each other. In reality, the two groups are not totally independent of each other as there are variables that are slightly correlated between the two groups, such as GDP and urbanisation rate. However, on the whole, it is true that the variables in the first group are rather opposite to those in the second group, so there is some independence between the two groups. This may be useful for modelling purposes as it suggests that there may be two different mechanisms underlying the variables in each group.

Now that we have analysed the statistical relationships between our crime variables, let us turn to the analysis of the individuals in our sample. In the graph below, we have plotted the representational qualities of each individual on our two axes:

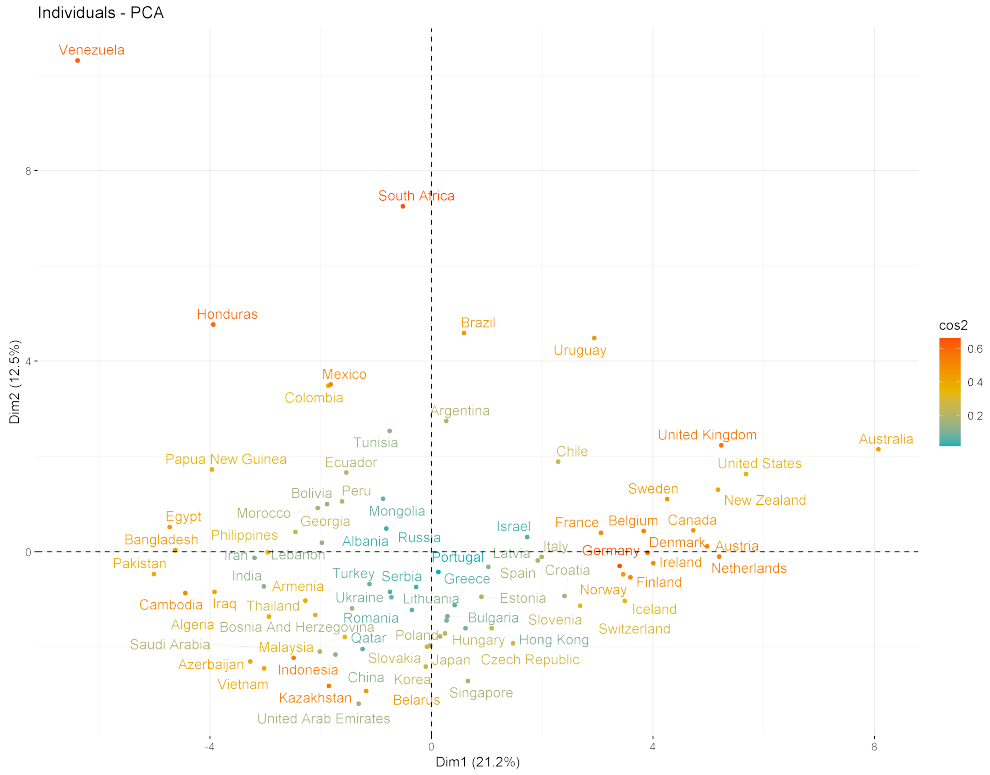

Figure 2.20:

This graph shows the countries that stand out in our sample. Indeed, Venezuela, South Africa, Honduras and Brazil have quite uniquely high statistics. A similar isolation on the other axis can be seen for Australia, the UK and the US. This confirms the analyses we made earlier and underlines the singularity of these countries.

Indeed, the analysis of the data allows us to highlight certain countries that stand out from the others. Venezuela, South Africa, Honduras and Brazil have particularly high statistics in terms of crime and associated socio-economic variables.

On the other hand, Australia, the United Kingdom and the United States appear isolated on the other axis, which may suggest that they have particular characteristics in terms of crime and socio-economic variables that differentiate them from other countries in the sample.

We will now see if this analysis is also reflected in the clusters we can create.

2.3.2 Clustering

Again, our data do not allow us to use the hierarchical clustering method. The density clustering method also proved ineffective.

We will now apply the k-means method. To do this, we need to establish the number of clusters we want.

To this end, we will again apply the elbow method, but this time apply it to the intra-cluster distance.

The graph below represents this distance, as a function of the number of clusters:

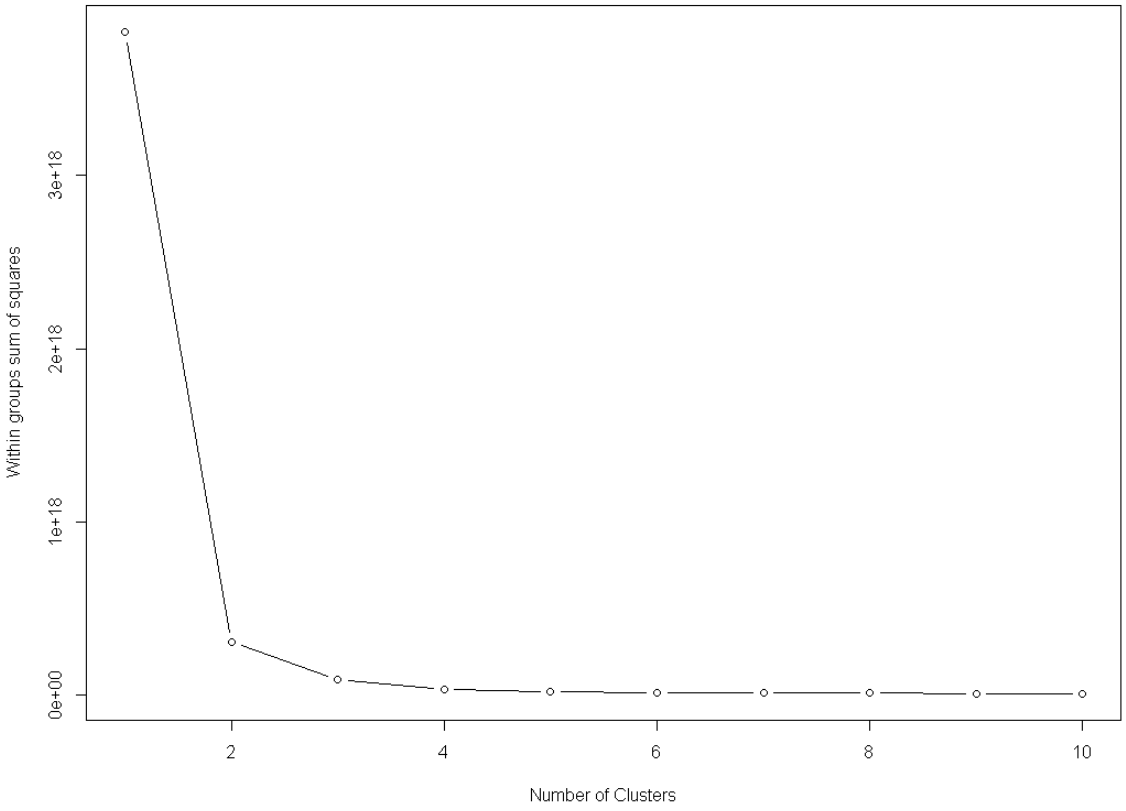

Figure 2.21:

Using the elbow method, we can conclude that this sample of countries can be grouped into 2 clusters. We can determine these with the k-means method.

This gives us the following representation:

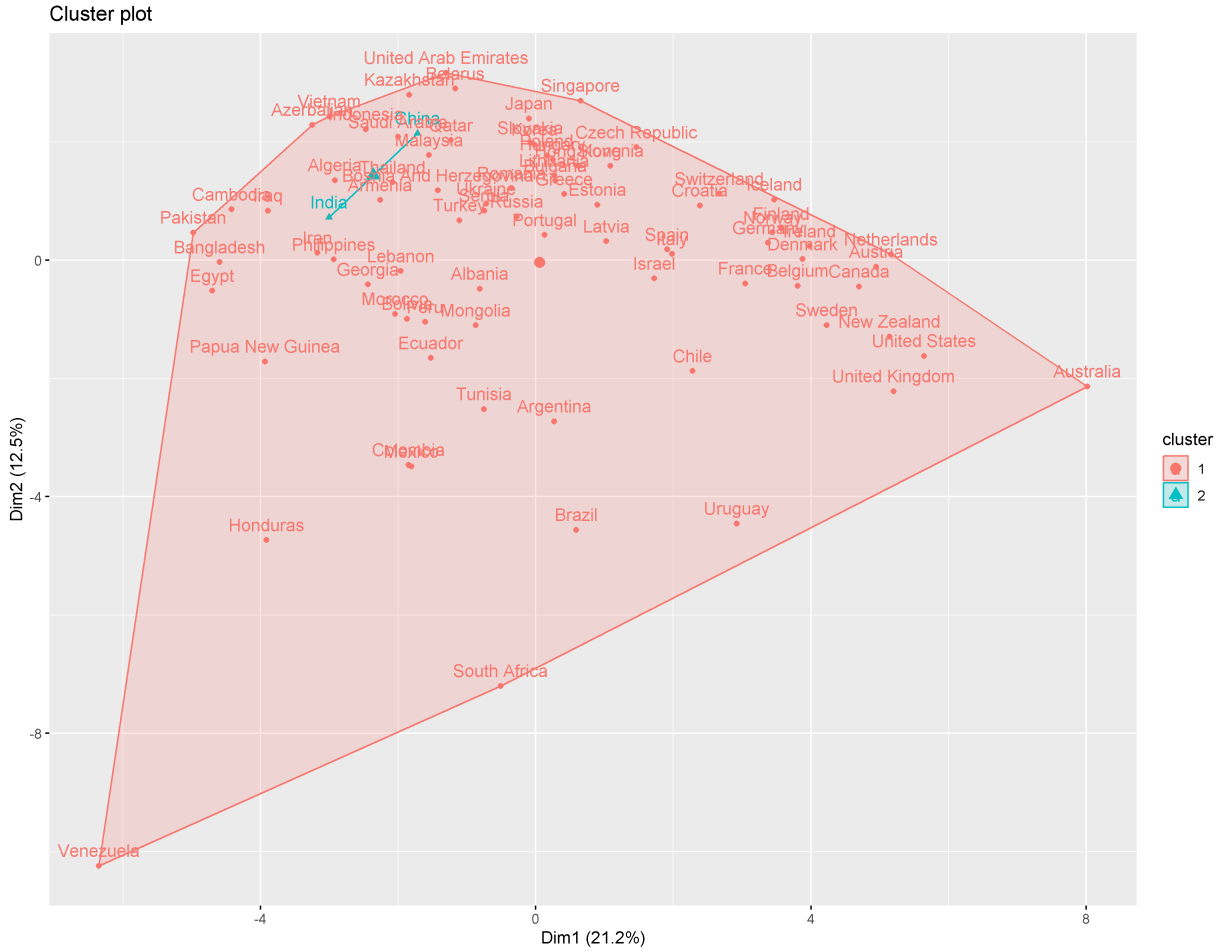

Figure 2.22:

We obtain exactly the same clusters as those obtained previously. So there is nothing more to add. This still confirms the quality of our sample.

Let us now see how these groups are characterised, using our socio-economic and crime variables:

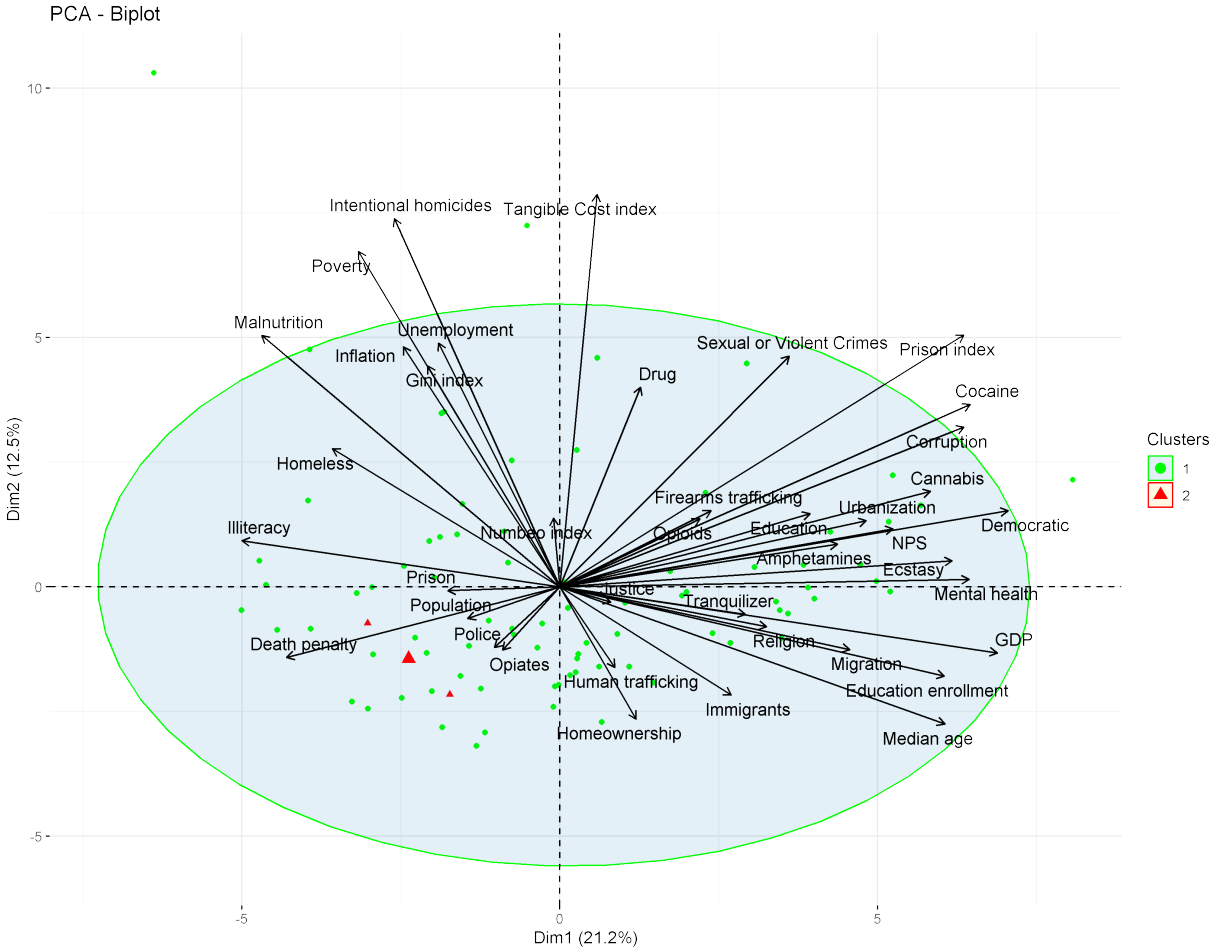

Figure 2.23:

This last graph reassures us about the choices of countries. The two slightly different countries, China and India, are different only in terms of population and the death penalty.

In conclusion, our data analysis revealed some relationships between socio-economic variables and crime statistics. We were able to identify two distinct groups of variables, one composed of malnutrition, inflation, unemployment, poverty, intentional homicide and the tangible cost index, and the other composed of violent and sexual crime, the prison population index, drug use, economic crime, the democratic index, the rate of urbanisation, the number of mental health practitioners, GDP, the number of students in higher education and median age. We also noticed that some countries stand out in our sample, such as Venezuela, South Africa, Honduras and Brazil, which have very high statistics, while Australia, the UK and the US are isolated on the other axis.

We were able to identify two distinct groups of variables that appear to be independent of each other, with countries that stand out in each of these groups.

On the one hand, we have identified a group of variables that are strongly correlated with poverty and violence, including intentional homicide. On the other hand, we identified a group of variables that are strongly correlated with drug use, economic crime and overall quality of life.

These results suggest that there is a significant relationship between poverty and violence in the countries studied, while drug use appears to be an important factor in understanding economic crime and overall quality of life.

These results are useful for better understanding the relationship between socio-economic variables and crime statistics. However, it is important to note that these correlations do not allow us to determine causal relationships between the variables. We hope that these results can be used in our future econometric modelling to identify factors that influence crime and to propose public policies to reduce it.

Chapter 3

Econometric models

Crime analysis is a complex and multidimensional subject that requires a methodical and rigorous approach. The crime regressions that follow aim to establish statistical relationships between the different explanatory variables and the different crime indices by country.

Among the explanatory variables considered in this analysis are Net Migration, Religion, Justice, Police, Education, Prison, Death Penalty, Unemployment Rate, Poverty, Homeless, Malnutrition, Mental Health, Drugs, Democracy, Homeowners, Students, Illiterates, Gini Index, GDP, Inflation, Median Age, Urbanization Rate and Immigrants.

Using statistical modeling techniques, these regressions can help understand the factors that contribute to crime in different social and economic contexts and inform public policies to prevent and combat it.

Our starting question for this study was the following: which of our different factors most influences our three crime indexes ?

We have therefore put as explanatory variables our different indexes and as explanatory variables the different empirical causes found in the literature to explain crime such as the percentage of GDP allocated to prisons, to justice, to education or if the death penalty is in force in the country or not with a dichotomous variable.

Then we thought of adding the level of unemployment in the country or factors related to poverty such as malnutrition or the number of homeless people registered, as we explained earlier in our study.

We first started with a simple linear model without logarithmic variables. And we have, as proven by the next regressions, satisfactory results with an R^2 around 65%. We then decided to test if there was a logarithmic relationship between our variables, hence the choice of log linearizing the variables that did not present multicollinearity.

Source	SS	df	MS
Model	13028.1674	23	566.44206
Residual	5378.71922	58	92.7365383
Total	18406.8866	81	227.245514

Number of obs	82
F(23, 58)	6.11
Prob > F	0.0000
R-squared	0.7078
Adj R-squared	0.5919
Root MSE	9.63

IndiceCri~eNumbeo	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SoldeMigratoire	.0000171	6.42e-06	2.67	0.010	4.26e-06	.00003
Religion	1.175306	1.128734	1.04	0.302	-1.084101	3.434714
Justice	-.4314316	2.374549	-0.18	0.856	-5.184609	4.321745
Police	-1.582423	1.594902	-0.99	0.325	-4.774966	1.61012
Education	1.084823	.8761168	1.24	0.221	-.6689156	2.838561
Prison	-1.067946	1.373146	-0.78	0.440	-3.816597	1.680704
Peinedemort	-2.063596	3.971903	-0.52	0.605	-10.01422	5.887031
Tauxdechomage	.0740929	.2938734	0.25	0.802	-.5141587	.6623445
Pauvrete	.274094	.1895425	1.45	0.154	-.1053166	.6535047
SDF	-.7655409	.5843816	-1.31	0.195	-1.935308	.404226
Malnutrition	.0757267	.1504309	0.50	0.617	-.2253934	.3768469
santementale	.0192129	.0248434	0.77	0.442	-.0305165	.0689423
Drogue	.4171986	.9419142	0.44	0.659	-1.468248	2.302645
Democratique	.6449479	1.239671	0.52	0.605	-1.836523	3.126419
Proprio	.0799869	.0800123	1.00	0.322	-.0801752	.240149
Etudiants	.0130958	.0803717	0.16	0.871	-.1477856	.1739771
Illetres	.5057439	.2693781	1.88	0.065	-.033475	1.044963
IndicedeGini	.1242454	.2163284	0.57	0.568	-.3087829	.5572737
PIB	-.0001931	.0001297	-1.49	0.142	-.0004527	.0000665
Inflation	.1500245	.074056	2.03	0.047	.0017853	.2982636
Agemedian	-.9829123	.3158155	-3.11	0.003	-1.615086	-.3507389
Tauxdurbanisation	-.0105183	.1260279	-0.08	0.934	-.2627906	.2417541
Immigres	-.0771054	.1662271	-0.46	0.644	-.4098451	.2556342
_cons	51.44341	18.81276	2.73	0.008	13.78559	89.10123

Figure 3.1: Results of the regression with Numbeo Index

Source	SS	df	MS
Model	4302.01653	23	187.044197
Residual	2235.94944	58	38.5508524
Total	6537.96596	81	80.7156292

Number of obs	82
F(23, 58)	4.85
Prob > F	0.0000
R-squared	0.6580
Adj R-squared	0.5224
Root MSE	6.2089

IndexCriminalit~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SoldeMigratoire	6.88e-06	4.14e-06	1.66	0.102	-1.40e-06	.0000152
Religion	.0329888	.7277518	0.05	0.964	-1.423765	1.489742
Justice	-.1822403	1.530991	-0.12	0.906	-3.246853	2.882372
Police	-.4752648	1.028314	-0.46	0.646	-2.533658	1.583128
Education	.3034191	.5648765	0.54	0.593	-.8273041	1.434142
Prison	-1.635966	.8853364	-1.85	0.070	-3.40816	.136227
Peinedemort	1.683726	2.560885	0.66	0.513	-3.442443	6.809895
Tauxdechomage	.1980127	.189475	1.05	0.300	-.1812627	.577288
Pauvrete	.3806198	.1222076	3.11	0.003	.1359947	.6252449
SDF	-.52827	.3767802	-1.40	0.166	-1.282478	.2259376
Malnutrition	.1329781	.0969904	1.37	0.176	-.0611692	.3271254
santementale	.02094	.0160178	1.31	0.196	-.011123	.0530031
Drogue	.7753101	.6072995	1.28	0.207	-.4403319	1.990952
Democratique	.5464588	.7992781	0.68	0.497	-1.05347	2.146388
Proprio	.0821722	.051588	1.59	0.117	-.0210923	.1854368
Etudiants	-.0379507	.0518196	-0.73	0.467	-.141679	.0657776
Illetres	-.0225832	.1736816	-0.13	0.897	-.3702447	.3250783
IndicedeGini	.149978	.1394778	1.08	0.287	-.1292172	.4291731
PIB	-.0000309	.0000836	-0.37	0.713	-.0001983	.0001365
Inflation	-.0082807	.0477476	-0.17	0.863	-.1038579	.0872966
Agemedian	-.2746611	.2036221	-1.35	0.183	-.6822551	.1329328
Tauxdurbanisation	.1124184	.0812566	1.38	0.172	-.0502343	.275071
Immigres	-.0341101	.107175	-0.32	0.751	-.2486441	.1804239
_cons	33.67795	12.12953	2.78	0.007	9.398056	57.95784

Figure 3.2: Results of the regression with Tangible Cost Index

Source	SS	df	MS
Model	7913.55282	23	344.067514
Residual	4755.51023	58	81.9915557
Total	12669.063	81	156.408186

Number of obs = 82
F(23, 58) = 4.20
Prob > F = 0.0000
R-squared = 0.6246
Adj R-squared = 0.4758
Root MSE = 9.0549

IndiceCrim~prison	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SoldeMigratoire	6.76e-06	6.03e-06	1.12	0.267	-5.32e-06	.0000188
Religion	.7879695	1.061331	0.74	0.461	-1.336515	2.912454
Justice	.40606	2.232751	0.18	0.856	-4.063277	4.875397
Police	.597457	1.499661	0.40	0.692	-2.404441	3.599355
Education	1.740754	.8237987	2.11	0.039	.0917417	3.389766
Prison	-.8262213	1.291148	-0.64	0.525	-3.410734	1.758291
Peinedemort	-.6405885	3.734717	-0.17	0.864	-8.116437	6.83526
Tauxdechomage	.5361088	.2763245	1.94	0.057	-.0170148	1.089232
Pauvrete	.0862687	.1782238	0.48	0.630	-.2704851	.4430225
SDF	-1.057642	.5494848	-1.92	0.059	-2.157555	.0422715
Malnutrition	.1404478	.1414478	0.99	0.325	-.1426907	.4235864
santementale	.0418286	.0233598	1.79	0.079	-.0049311	.0885884
Drogue	.8860781	.885667	1.00	0.321	-.886777	2.658933
Democratique	1.552792	1.165643	1.33	0.188	-.780496	3.886079
Proprio	-.0595768	.0752343	-0.79	0.432	-.2101747	.0910211
Etudiants	-.0462631	.0755722	-0.61	0.543	-.1975373	.105011
Illetres	.0417329	.253292	0.16	0.870	-.465286	.5487519
IndicedeGini	.1151445	.2034102	0.57	0.574	-.2920252	.5223142
PIB	.000075	.000122	0.61	0.541	-.0001692	.0003191
Inflation	-.0469062	.0696337	-0.67	0.503	-.1862931	.0924807
Agemedian	-.0519889	.2969563	-0.18	0.862	-.6464115	.5424337
Tauxdurbanisation	.0779668	.1185021	0.66	0.513	-.1592409	.3151745
Immigres	-.1349517	.1563007	-0.86	0.391	-.4478215	.1779181
_cons	25.1448	17.68934	1.42	0.161	-10.26425	60.55386

Figure 3.3: Results of the regression with Prison Index

3.1 Regression with Log Transformation

When attempting to log-linearize variables in a regression model that exhibits multicollinearity, we may exacerbate this problem. Log-linearization involves taking the logarithm of some of the independent variables, which can have the effect of reducing the correlation between these variables and the dependent variable but can also increase the correlation between these independent variables.

In this case, if you log-linearize variables in a model that exhibits multicollinearity, it can reinforce the correlation between the independent variables and accentuate the problem of multicollinearity, which can make the results of the regression model even less reliable.

Therefore, when working with a model that exhibits multicollinearity, it is preferable to take steps to reduce the correlation between the independent variables rather than attempting to log-linearize them. This may involve removing certain variables, grouping variables into factors, or using estimation techniques that are less sensitive to multicollinearity, such as ridge regression or Lasso regression.

Source	SS	df	MS
Model	10000.521	25	400.02084
Residual	955.386188	19	50.2834836
Total	10955.9072	44	248.997891

Number of obs = 45
F(25, 19) = 7.96
Prob > F = 0.0000
R-squared = 0.9128
Adj R-squared = 0.7981
Root MSE = 7.0911

IndiceCriminali~o	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnSoldeMigratoire	1.509653	1.535546	0.98	0.338	-1.704282	4.723588
lnJustice	4.984625	5.447648	0.92	0.372	-6.417433	16.38668
lnPrison	-2.059871	1.903085	-1.08	0.293	-6.043073	1.923331
lnTauxdechomage	-1.868877	3.10892	-0.60	0.555	-8.375921	4.638167
lnSDF	.0107497	1.450701	0.01	0.994	-3.025602	3.047102
lnMalnutrition	5.416219	3.739958	1.45	0.164	-2.411604	13.24404
lnsantementale	.9640837	2.085587	0.46	0.649	-3.401101	5.329268
lnIlletres	-.0239904	2.39124	-0.01	0.992	-5.028913	4.980932
lnPopulation	6.426481	2.131374	3.02	0.007	1.965463	10.8875
Religion	-.044403	1.452132	-0.03	0.976	-3.083749	2.994943
Justice	-7.696055	6.816415	-1.13	0.273	-21.96297	6.570865
Police	-3.680995	1.718474	-2.14	0.045	-7.277802	-.0841875
Education	1.243968	1.408096	0.88	0.388	-1.703211	4.191147
Peinedemort	-12.31656	5.467633	-2.25	0.036	-23.76045	-.872676
Pauvrete	.1604005	.3145425	0.51	0.616	-.4979446	.8187456
Drogue	-.8508791	1.087108	-0.78	0.443	-3.126223	1.424465
Democratique	1.582495	1.884024	0.84	0.411	-2.360812	5.525802
Proprio	.1877142	.1403871	1.34	0.197	-.1061193	.4815478
Etudiants	-.1432878	.1402495	-1.02	0.320	-.4368333	.1502578
IndicedeGini	-.1920992	.4025328	-0.48	0.639	-1.03461	.6504117
PIB	-.0001363	.0001695	-0.80	0.431	-.0004909	.0002184
Inflation	-.0914488	.4694529	-0.19	0.848	-1.074025	.8911273
Agemedian	-1.9518	.6965299	-2.80	0.011	-3.409654	-.4939466
Tauxdurbanisation	-.0805819	.1630163	-0.49	0.627	-.421779	.2606152
Immigres	.2210133	.2833373	0.78	0.445	-.3720185	.8140451
_cons	-14.68923	30.73172	-0.48	0.638	-79.01147	49.63301

Figure 3.4: Results of the regression with Log-linearisation

3.1.1 Regression results

We can begin by noting our R^2 value, which is 0.9128. This means that our model choice explains 91.28% of the variance in the variable we are trying to explain, in this case, the Numbeo crime index. If we are in the presence of multicollinearity, we can observe these problems:

- the values/signatures of the coefficients are contradictory: they do not agree with the knowledge of the domain,
- high variance of the estimated parameters and confidence intervals around the important parameters
- Student's t tests are not very significant although the explanatory variables are jointly highly significant
- despite this, the coefficient of determination can be very high
- instability of the results (addition or deletion of some observations leads to strong changes)

3.1.2 Statistical Tests

Let's now try the multicollinearity test. To do this, we may use the estat vif command with the collin command.

Variable	VIF	1/VIF
lnJustice	23.45	0.042653
Justice	19.76	0.050608
Agemedian	19.01	0.052598
PIB	14.04	0.071207
Immigres	13.49	0.074122
Democratique	11.82	0.084587
IndicedeGini	11.73	0.085245
lnsantemen~e	9.27	0.107899
Pauvrete	8.56	0.116759
Etudiants	8.41	0.118949
lnPopulation	7.72	0.129601
lnIlletres	7.03	0.142282
lnMalnutri~n	6.99	0.142968
lnSDF	6.69	0.149394
Tauxdurban~n	6.01	0.166498
lnTauxdech~e	5.67	0.176378
Peinedemort	5.50	0.181947
lnSoldeMig~e	5.48	0.182508
Proprio	5.37	0.186189
Inflation	4.79	0.208575
lnPrison	4.49	0.222585
Education	3.70	0.270117
Religion	3.27	0.305716
Drogue	2.45	0.407896
Police	2.30	0.433969
Mean VIF	8.68	

Figure 3.5:

We now have with this command, the variance inflation factors (VIFs) for the independent variables in our model. It calculates the ratio of variance when you have multiple independent variables.

By convention a variable whose VIF values are greater than 10 may merit further investigation, the tolerance is defined as 1/VIF, is used by many researchers to check on the degree of collinearity so a tolerance value lower than 0,1 is comparable to a VIF of 10.

This measure means that the variable can be considered as a linear combination of other independent variables. Based on these high vif values, we decided to observe the vif of the model without log linearization. This clearly gives a better result for multicollinearity, so it is better to keep our model without log linearization.

3.2 First Model with Numbeo Index

In this part, we will this discuss the results of this graph 3. The endogeneity tests having been performed, we did not find any endogenous independent variables in our model.

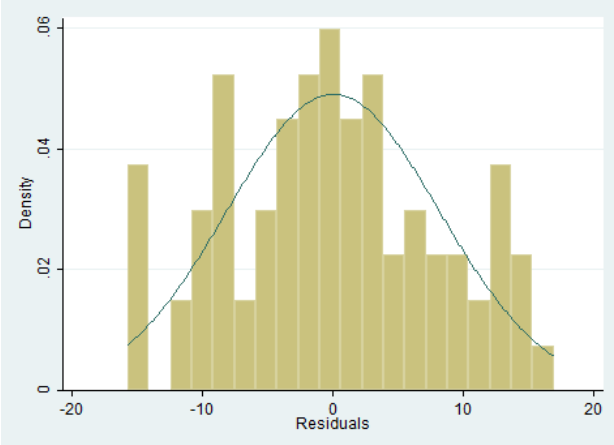

Figure 3.6:

Upon first inspection to the resulst, the residuals of our model appear to follow a normal distribution, and that is what we will verify using the command :

Skewness/Kurtosis tests for Normality

				joint	
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
resid3	82	0.7172	0.0760	3.39	0.1833

Figure 3.7:

This command is used to perform a test for skewness and kurtosis on the residuals of a regression model. Skewness refers to the degree of asymmetry in the distribution of residuals, while kurtosis measures the degree of peakedness or flatness in the distribution.

It conducts a test to determine whether they are significantly different from those expected under a normal distribution. This help us assess whether the assumption of normality of residuals in a regression model is violated.

Here, the value of our test statistic is 3.39. If the value is less than 0.05, we can doubt the normality of the residuals. In our case, our residuals follow a normal distribution. We will now perform tests to observe potential correlations using the Cameron-Trivedi test.

3.2.1 Cameron Trivedi Test

This test allows us to determine whether the model data is correlated or uncorrelated with measurement errors, and whether there is a normal distribution of errors.

```
Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	82.00	81	0.4480
Skewness	15.28	23	0.8845
Kurtosis	3.83	1	0.0503
Total	101.11	105	0.5894

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of IndiceCriminaliteNumbeo

         chi2(1)      =     0.07
         Prob > chi2  =   0.7985
```

Figure 3.8:

We have here the results of the White test and the Cameron and Trivedi test. Since all the p-values are greater than 5%, we do not reject the null hypothesis of the Cameron-Trivedi test, which states that the model errors are independent of each other, and therefore there are no sequential correlations between the errors.

When the variance of a model's errors is not the same for all observations, it is referred to as heteroscedasticity. The presence of this phenomenon is problematic because we rely on an OLS model that assumes homoscedasticity. Indeed, the causes of heteroscedasticity can be numerous:

- When errors are related to the values taken by an explanatory variable
- When observations represent averages (which is why the "Moyennedesindex" variable is not taken into account in our regressions)
- When using grouping into categories

We were able to verify using the sktest command in Stata that our results followed a normal distribution and will now proceed with the Breusch-Pagan test to determine whether or not we have heteroscedasticity.

3.2.2 Breusch-Pagan Test

For this test, the null hypothesis is that the variance of errors in this regression is constant for all values of the independent variable. In other words, there is no heteroscedasticity (meaning there is homoscedasticity) in the model.

In our case, the test statistic is 18.35, which should be compared to the critical value of a chi-squared distribution with 25 degrees of freedom at 5%, which is 41.37. Since the critical value is greater than our test statistic, we cannot reject H0. Additionally, observing the p-value (0.8628) indicates that it is greater than our predetermined threshold of 5%. Therefore, we are not dealing with a problem of heteroscedasticity.

```
. estat hettest, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: SoldeMigratoire Religion Justice Police Education Prison Peinedemort Tauxdechomage Pauvrete SDF
                    Malnutrition santementale Drogue Democratique Proprio Etudiants Illetres IndicedeGini PIB Inflation
                    Agemedian Tauxdurbanisation Immigres

         chi2(23)     =    15.11
         Prob > chi2  =   0.8905
```

Figure 3.9:

This test is looking for heteroskedasticity linked to all independent variables, rather than the fitted values of the dependent variable. With a p-value of 0,8905 the test indicates there is no significant heteroskedasticity related to our independent variables. Nonetheless this is a test that looks at all independent variables together? By testing every variables individually you can end up with a situation where homoskedasticity is masking heteroskedasticity.

When there is model with not much independent variables it is possible to do it individually. In our case, we can run a multiple hypotheses test by specifying a multiple-testing adjustment in addition to the rhs option. By doing this, the probability of incorrectly reject the null hypothesis (make a false positive or 1 error) increases xith each additional test. We need to adjust the p-values of every test to take this to account. For adjustment we can name the Bonferroni correction or the Sidak correction.

Since our model has no time series, it is not necessary to check for autocorrelation of errors.

By keeping our first model of regression with our R^2 at 70,78% we can repeat what we previously did with the different test :

```
. estat imtest

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	82.00	81	0.4480
Skewness	15.28	23	0.8845
Kurtosis	3.83	1	0.0503
Total	101.11	105	0.5894

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of IndiceCriminaliteNumbeo

         chi2(1)      =     0.07
         Prob > chi2  =   0.7985
```

Figure 3.10:

We now have the White's test and the Breusch-Pagan test with the same null hypothesis which is that the variance of the residuals is homogenous. On that account, if the pvalue is very small we would have to reject the hypothesis that the variance is homogenous, so for this case, the variance is thus not homogenous.

The calculated test statistic is 0,07. She needs to be compared to the critical chi-square value at 81 degrees of freedom at the 5% significance level is approximately 108,3. So we do not reject the null hypothesis. Moreover, the p-value of the test (0.7985) is greater than 5%. We are therefore not in the presence of a heteroscedasticity problem.

3.2.3 Fisher Test and Significativity of the regression

The Fisher test looks for the existence of at least one significant explanatory variable with the following assumptions :

- H0 : all coefficients of the model are zero
- H1 : there is at least one non-zero coefficient

The arbitration is done by comparing the value of the estimated F-statistic to the one tabulated by Fisher. The stata software automatically provides the probability associated with the calculated F-statistic, which greatly facilitates the analysis. It is therefore sufficient to compare the probability associated with the F-statistic with the 5% threshold.

If the probability associated with the calculated F-statistic is less than 5%, then the H0 hypothesis will be rejected in favor of the alternative hypothesis that the regression is globally significant.

In our case, the Fisher statistic is F= 6.11 and the associated probability is less than 5% (0.000 $<$ 0.05), but the statistic read from the Fisher table at 23 and 58 degrees of freedom at the 5% threshold is 1.70: therefore, the null hypothesis is rejected and the model is globally significant. This result is consistent with the value of the adjusted R2 statistic and confirms the accuracy and relevance of our model.

The model is globally significant and has a good quality.

Let's now move on to the significance of each variable and the Student's t tests with :

$$H_0 : \beta_i = 0 \tag{3.1}$$

$$H_1 : \beta_i \neq 0 \tag{3.2}$$

The Student's t test consists in comparing the value obtained thanks to the following formula:

$$ti = \beta i / \sigma i \tag{3.3}$$

where sigmai is the standard deviation of the studied coefficient, with the value of the associated Student table. If ti> t0,25 we reject H0 and the coefficient is significantly different from 0.

t0,25 = 1.96 according to Student Table, on our stata output by taking the "t" column and by taking the absolute value we can observe that the only significant variables are "SoldeMigratoire", "AgeMedian" and the constant.

We can verify this by looking at the pvalues by seeing if they are lower than 0,05 or by looking at the confidence range by judging if it contains 0 and then being non significant otherwise, if the interval does not contains 0 the variable is significant.

The estimation results show that only three variables are statistically significant at the 5% level.

3.2.4 Interpretation of the significant values

Since we have a negative sign associated with our significant variable median age, we can tell that the median age of the population is inversely related to the crime index.

In our case, the coefficient of the median age is -0.9829123 which means that for each increase of one year in the median age of the population, the Numbeo crime index decreases on average by 0.9829123 units, all else being equal. We also have a positive relationship between the net migration variable and our variable to be explained which is the Numbeo crime index.

Since the coefficient of net migration is 0.0000171, for each increase of one unit in net migration, the crime index increases on average by 0.0000171 units, all else being equal.

This can be explained by the fact that migrant populations are more likely to be in disadvantaged economic and social situations. This may make them more vulnerable to financial pressures and the temptation to commit crimes. The constant being the last significant variable of the regression.

This means that even if all the independent variables in the regression are zero, there remains a predicted value for the crime index that is different from zero. We can explain this as omitting some factors that would help explain the crime index but we cannot quantify the impact of these unmeasured factors.

We can therefore observe that migration can cause ethnical tension and also insecurity, which is measured by the Numbeo Index.

These results can be qualified by the relativity of the validity of this chosen index. We would like to remind you that this index is based on feelings and that the results must be observed with hindsight.

That's why we wanted to add and calculate ourselves other indexes to compare more quantitative indexes with the first index collected based on the feeling. This is what we will see with the next regressions.

3.3 Second Model with the index based on tangible cost

In this part, we will this discuss the results of this graph 3. It is important to specify that all the tests carried out previously have been reproduced for this and the next regressions but have not been submitted for the sake of concision and length of the study.
For this regression with the index based on tangible costs as the variable to be explained, at the 5% threshold, only two variables are significant, namely, the poverty variable and the constant of the regression.

The explanation of the creation of this index having been made earlier in the file we can try to interpret the two significant variables. The constant being significant for the same reasons as in the previous regression of the Numbeo crime index, we can move to the coefficient of poverty.

An attempt to explain this would be that areas with a higher poverty rate would tend to have higher crime rates due to factors such as lack of resources and economic opportunities, increased social tension and vulnerability to committing "petty" crimes.

For every one unit increase in the poverty rate, the tangible cost-based crime index increases by an average of 0.3806198 units all else being equal.

It would be interesting to move to a 10% threshold for this regression rather than 5% because the prison variable would become significant. Indeed, there is an inverse relationship between the propensity of resources allocated to prisons and the crime index based on the tangible cost of crime.

Here, for every one unit increase in the percentage of GDP allocated to prisons, the crime index associated with the tangible costs of crime decreases on average by 1.635966 units, all else being equal. We could explain this as an indication that investments in crime prevention programs and alternatives to imprisonment can reduce crime.

3.4 Third Model with the index based on legal sentences

In this part, we will this discuss the results of this graph 3. For the next model we are going to regress the last index of criminality that we made, based on legal sentences.

Even though our R^2 are increasing for each regression, we can observe that the R^2 adjusted is decreasing which is a better indicator of relevance of our model.

This time education is the only variable that is significant at the 5% level, in fact for each increase of one unit of education level, the crime index based on judicial convictions increases by an average of 1.740754 units, all else being equal.

A possible interpretation would be that higher levels of education may be associated with exposure to different form of crime, especially tax fraud or corruption. In the massiv educated countries, there is enormous investments in statistics instistution, which leads to a greater recording of crime. Therefore the statistic correlation between investment in education and crime might not be a causal relation.

Again, it would be interesting to move to a 10% threshold to see that the variables of homelessness, health status, and unemployment rate become significant.

For the coefficient of the number of homeless, a curious relationship is observed: a higher level of homelessness may be associated with a decrease in crime. A possible explanation is that the public authorities very rarely consider this crime and would participate in making it invisible. This would statistically decrease crime, but only because it would be ignored.

For the mentalhealth variable, a positive relationship is observed, i.e., for each increase of one unit of caregivers, the crime index based on court convictions increases by an average of 0.0418286 units, all else equal.

We could try to explain this phenom by a kind of bias that would be created, where countries with a large number of mental health practitioners would just be a reflection of a rich and developed country that would take better care of its population thanks to more knowledgeable professionals on the subject and better documentation. Like the education, this could be only a correlation between mental health and crime.

The last variable that is significant at the 10% level is the unemployment rate. Indeed, a one unit increase in the unemployment rate would lead to a 0.5361088 unit increase in the crime index based on judicial convictions, all other things being equal.

Attempted explanations include the loss of a job, which would change an individual's life and increase the unemployment rate, which, in a precarious situation, would push the individual to commit crimes in order to regain the standard of living that he or she had when active. The link between these two effects is well documented and we will see it in more detail in the next and final chapter.

We have seen throughout this chapter the variables that had a significant impact on our different indices. But among our numerous variables, those that we suspected of having a significant impact on our indexes turned out not to have any.

Indeed, to quote the variable death penalty, it is often justified as a disincentive to commit crimes and misdemeanours given the possible penalty. It is constantly questioned, especially in the United States where it varies from state to state. It has been proven, according to our regressions, that it has no impact on crime.

We can also cite the so-called "precariousness" factors that were thought to have an impact on crime, such as poverty, malnutrition or the number of illiterates, which again did not impact our crime index variables. These results must be qualified, of course, because even if we did not find a significant effect, this may be due to a measurement error during the collection of the data, or it may be correlated with another variable that is significant.

We also found that there was a statistical indirect relationship between wealth and crime through the unemployment rate. If an individual's wealth level drops, he or she may be led to commit crimes to regain a financial situation that allows him or her to live.

This relationship is interesting, and we will therefore in the next chapter find literature that explains this relationship and legitimizes the variables that we found significant in our regressions.

After having found through regressions the factors that influenced our different indices, we can confirm our choice to build new indicators to complement the one provided by Numbeo, which remains to be qualified by its construction based more on feelings than on costs and tangible penalties.

It is important to bring several indicators to support statements so as not to draw conclusions too quickly and to step back and take into account the socio-economic situations of the countries and the various crises that can occur and alter our different values, hence our choice of the year 2018 to collect our data.

We will now consult some literature that confirms the interpretations we have made.

Chapter 4

Validation of the interpretations of our econometric models using the academic literature

During our study, we noticed that very few articles and books were structured in the same way as our study.

Indeed, there is very little literature on econometric studies that has chosen a panel as large as the one we have chosen. The studies generally focus on the crime of one country or a group of countries, as the subject is so full of topics to be treated.

For this last part, we will search for literature that supports our regression results, namely: median age, prison investment, unemployment rate, education, and homelessness and whether their relationship with crime has already been studied.

To begin with, intuitively, we can establish a relationship with our first variable, median age, and crime: as the population ages, therefore the level of wealth increases and therefore we think of a decrease in the need to commit crimes, because the link between poverty and crime is empirically obvious.

We will take to support this link an article published in 2002 entitled "The dynamic of poverty and crime"[20] by Haiyun, Zhao, Zhilan, Feng, Carlos, Castillo-Chavez.

They explain that poverty is an important factor that contributes to crime, but the relationship between the two is complex and multifactorial. They implement mathematical models to better understand the dynamics of poverty and crime and to identify the key factors that influence this relationship. They highlight that poverty can lead to crime through various channels, such as lack of economic opportunity, social frustration and isolation, and the influence of criminal peers.

They conclude that policies aimed at reducing poverty can help reduce crime by improving economic and social opportunities for the disadvantaged. Policies that link the two aspects are more effective than those that focus on only one of the two factors.

Another argument for the relationship between median age and crime would be that, as many studies have recently shown, poorer people die faster than wealthier people as explained in these two articles "Social disadvantage, economic inequality, and life expectancy in nine Indian states"[4] by Sangita Vyas, Payal Hathi, Aashish Gupta and "The impact of poverty and deprivation at the end of life: a critical review"[17] by Jane Rowley , Naomi Richards, Emma Carduff and Merryn Gott.

The literature on the link between poverty and crime is very rich and well documented, which is why the addition of another study, this time on Europe, seemed relevant, so we choose "Poverty and Crime in Europe: An Empirical Analysis with Panel Data (2000 – 2008)"[10] by M. Georgiou.

Here, author Miltiade N Georgiou uses panel data on poverty and crime for 27 European countries over the period 2000-2008. He applies econometric models to study the relationship between these two variables, considering factors such as unemployment, economic growth, and public spending.

His work shows that there is a positive correlation between poverty and crime in Europe, but that this relationship is stronger for countries with high levels of relative poverty and unemployment. Furthermore, he finds that public spending has a negative effect on crime, suggesting that public policies can play an important role in reducing crime.

Finally, he emphasizes the importance of considering other socio-economic factors such as the level of human development and the degree of social protection in the country when analyzing the relationship between poverty and crime. Indeed, these factors may influence the way poverty affects crime.

These two studies justify at the same time the importance of the unemployment rate, which we found in our regressions, linked to crime and more precisely in our case in the construction of an index that evaluates it.

We will now try to find some literature that links the level of education and crime. At first sight, one thinks of a negative relationship between the level of education and the level of crime in a country; that is, the more educated a country is, the less crime there would be as discussed in the study "The effects of education on crime"[13] by W. Groot, H. M. van den Brink.

We will now try to find some literature that links the level of education and crime. At first sight, one thinks of a negative relationship between the level of education and the level of crime in a country; that is, the more educated a country is, the less crime there would be as discussed in the study "The effects of education on crime"[13] by W. Groot, H. M. van den Brink.

We found to contradict this a positive relationship in our regressions and tried to explain this by the fact that a person with a good level of education has consequently a better level of wealth and can therefore be linked to another type of crime: fraud. This crime is indeed less recorded, even more so in countries where corruption is very present. This could explain this relationship, which is statistical and not causal.

With the support of this article "Fraud reporting in Catalonia in the Internet era: Determinants and motives"[14] by Steven Kemp. This article highlights the importance of trust in institutions and regulations in encouraging fraud reporting. The results suggest that trust in regulators is a key factor influencing the decision to report or not report financial fraud. This confirms that fraud reporting is influenced by the country's democratic system and its level of corruption.

Moreover, the argument of the social situation of the accused is considered in correctional sentencing, as shown in the article "Educational Background and Sentencing Practice in Criminal Courts - An Empirical Study"[5] by A. Badó, E. Halász, E. Nagy must be taken into account.

They show that the educational level of defendants is significantly correlated with the severity of the sentence they receive. Defendants with higher levels of education receive lighter sentences on average than defendants with lower levels of education. The authors suggest that this may be due to implicit bias or stereotyping of those with higher levels of education. This would provide another explanation for the relationship between education and crime.

The last two relationships we wanted to explain with literature are the percentage of GDP allocated to prisons lowers crime and the increase in the number of homeless people lowers the level of crime. For the relationship of the prison variable on crime we chose to rely on the article "Allocating Resources Among Prisons And Social Programs In The Battle Against Crime"[7] by John J. Donohue III, Peter Siegelman.

In this article, the authors examine the financial and social costs associated with mass imprisonment, such as prison overcrowding, mental health costs, and the high rate of recidivism of offenders after their release. They suggest alternatives such as crime prevention programs, substance abuse treatment programs, education programs, and reentry programs for prisoners.

Therefore, better management of allocations to prisons and inmates, and the introduction of rehabilitation and support programs for inmates, reduces the likelihood of recidivism and thus lowers crime.

Finally, we will try to explain the negative relationship between the number of homeless and crime. Our regression showed us that an increase in the number of homeless people decreased crime. Indeed, Joshua T. Ellsworth's article "Street Crime Victimization Among Homeless Adults: A Review of the Literature"[8] showed that homeless people have a much higher risk of being victims of crime than the general population.

The most common types of crime were theft, physical assault, and sexual violence. The study also found that homeless women were particularly vulnerable to sexual violence.

However, despite the high levels of victimization, most homeless people did not report their assaults to the police. Reasons for underreporting included fear of retaliation, shame, and the perception that authorities would not be able to help them. This would explain this negative relationship.

We could mention that in our various literature searches, new factors that we had not considered may prove to be crucial for other articles and studies. We could mention the relationship of the media to crime or the culture of gangs, even if these are not causes of the same magnitude in the different countries we have chosen, it would be interesting to study them.

We also found interesting to study for potential another study, the crime that is not reported. Indeed, according to the CSEW[1] (Crime Survey for England and Wales) of the year 2020, only 41% of crimes were reported to the police.

The methods of measuring these under-reported crimes should therefore be taken into account. The reasons for a high level of "unreported crime" can indeed be diverse: a high level of corruption in the country, or public authorities considered too slow to resolve conflicts...

During our data collection, this factor came to mind, but the only way we could get a view on this corruption was to observe the democratic status of the country. In hindsight, we could see that this indicator of democracy was not sufficient to understand and consider the complexity of the countries.

It could be also be very interesting, as the United Nations did in the Global Study on Homicide, to study the different relationship of countries to firearms and how this impacts their crime rate, indeed an analysis of the legislation in force should have a direct relationship with crime in the country and it would be interesting to measure this.

Searching the literature to support econometric regression results is an important step in the research process. Econometric regression results should be interpreted with caution and supported by relevant prior research.

A literature review can help validate the econometric regression results by showing their consistency with findings from previous studies. In addition, it can help identify potential confounders or methodological limitations that may affect the econometric regression results.

Ultimately, a comprehensive literature search is essential to ensure the reliability and validity of the econometric regression results and to enhance the relevance of the research.

Conclusion

The data analysis conducted in the second chapter revealed correlations between socio-economic variables and crime statistics. The variables were grouped into two distinct categories, one linked to poverty and violence, and the other linked to drug use, economic crime and overall quality of life. Several countries stood out in the sample due to high crime statistics.

Three crime indices were created in the first chapter. Those were based on the total cost of crime to society, prison distribution, and a mixture of survey and Numbeo data. The limitations of the study were also discussed.

These findings provide a better understanding of the relationship between socio-economic variables and crime statistics, and may aid in future econometric modelling and policy proposals to reduce crime. The third and forth chapters provide a comprehensive review of various studies and articles that aim to understand the complex relationship between poverty, education, and crime.

The literature highlights the significance of poverty in driving criminal behavior, and the importance of policies that address economic and social opportunities for the disadvantaged.

Additionally, the negative correlation between education and crime is widely acknowledged, but the relationship is more complicated than originally thought, with fraud being a potential factor in the positive relationship found in some cases.

Furthermore, the allocation of resources to social programs and rehabilitation initiatives for prisoners can lower the crime rate by reducing recidivism.

Finally, the negative correlation between the number of homeless people and crime suggests that providing shelter and support for homeless individuals can decrease criminal behavior.

Overall, this literature review highlights the need for a multifaceted approach that considers economic, social, and institutional factors when addressing crime and its root causes.

Bibliography

[1] Crime in England and Wales - Office for National Statistics.

[2] dataUNODC |.

[3] Indice de Criminalité par Pays 2018.

[4] Social disadvantage, economic inequality, and life expectancy in nine Indian states | PNAS.

[5] Attila Badó, Edina Halász, and Erzsébet Nagy. Educational Background and Sentencing Practice in Criminal Courts – An Empirical Study. *Zeitschrift für die gesamte Strafrechtswissenschaft*, 128(4):1193–1206, April 2017. Publisher: De Gruyter.

[6] Jonathan P. Caulkins, Rosalie Liccardo Pacula, Susan Paddock, and James Chiesa. Social Costs of Drug Consumption. In *School-Based Drug Prevention*, What Kind of Drug Use Does It Prevent?, pages 95–112. RAND Corporation, 1 edition, 2002.

[7] John J. Donohue III and Peter Siegelman. Allocating Resources Among Prisons And Social Programs In The Battle Against Crime. *The Journal of Legal Studies*, 27(1):1–43, 1998. Publisher: [The University of Chicago Press, The University of Chicago Law School].

[8] Joshua T. Ellsworth. Street Crime Victimization Among Homeless Adults: A Review of the Literature. *Victims & Offenders*, 14(1):96–118, January 2019. Publisher: Routledge _eprint: https://doi.org/10.1080/15564886.2018.1547997.

[9] France. *Code pénal : annoté*. Codes Dalloz. Dalloz, Paris, 119e éd., 2022 edition, 2021. Country: FR 21 cm. En appendice, table chonologique. Notes bibliogr. Index.

[10] Miltiades N. Georgiou. Poverty and Crime in Europe: An Empirical Analysis with Panel Data (2000 – 2008), January 2011.

[11] B. Gompertz. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. January 1815.

[12] Statistics Canada Government of Canada. Crime severity index and weighted clearance rates, Canada, provinces, territories and Census Metropolitan Areas, April 2021. Last Modified: 2022-08-02.

[13] Wim Groot and H. Maassenvandenbrink. The effects of education on crime. *Applied Economics*, 42:279–289, February 2010.

[14] Steven Kemp. Fraud reporting in Catalonia in the Internet era: Determinants and motives. *European Journal of Criminology*, 19, July 2020.

[15] Cesare Lombroso, Mary Gibsonand, Nicole Hahn Rafter, and Mark Seymour. *Criminal Man*. Duke University Press, 2006.

[16] Kathryn E. McCollister, Michael T. French, and Hai Fang. The Cost of Crime to Society: New Crime-Specific Estimates for Policy and Program Evaluation. *Drug Alcohol Depend*, 108(1-2):98–109, April 2010.

[17] Jane Rowley, Naomi Richards, Emma Carduff, and Merryn Gott. The impact of poverty and deprivation at the end of life: a critical review. *Palliat?Care*, 15:26323524211033873, January 2021. Publisher: SAGE Publications Ltd STM.

[18] G. Tarde. Qu'est-Ce Que Le Crime? *Revue Philosophique de la France et de l'Étranger*, 46:337–355, 1898. Publisher: Presses Universitaires de France.

[19] Jean-Marie Tremblay. Étienne De Greeff, INTRODUCTION à LA CRIMINOLOGIE. VOLUME I, February 2005.

[20] Haiyun Zhao, Zhilan Feng, and Carlos Castillo-Chávez. The dynamics of poverty and crime. January 2002.