







ARTICLE

An application of the k -means clustering method to National Public Safety Data - the case of intentional homicides

 Henrique José de Paula Alves,^{*1}  Felipe Augusto Fernandes,²  Édipo Menezes da Silva,³ and  Ben Dêivide de Oliveira Batista⁴

¹Researcher at Institute of Applied Economic Research - IPEA, Rio de Janeiro - RJ, Brazil

²Researcher at State University of Campinas - UNICAMP, Campinas - SP, Brazil

³Researcher at Butantan Institute, São Paulo - SP, Brazil

⁴Federal University of Sao Joao del-Rei - UFSJ, Ouro Branco - MG, Brazil

*Corresponding author. Email: jpahenrique@gmail.com

(Received: October 9, 2023; Revised: June 6, 2024; Accepted: June 21, 2024; Published: February 11, 2025)

Abstract

The purpose of this article is to assist Brazilian authorities in directing resources to reduce the rate of intentional homicide per 100,000 inhabitants in the federative units of Brazil. In this sense, linked to the information made available by the National Secretariat for Public Security (Senasp), it was decided to apply the grouping technique known as k -means to group the FU's by similar rates. Three clusters were found and, then, a detailed description of each of the FU's belonging to each of the three clusters was carried out. We noticed that only the FU's that form cluster 1 have a homicide rate below 16 victims per 100,000 inhabitants, the limit imposed by the 2021 National Public Security Plan.. This plan was prepared by Brazilian Ministry of Justice, Senasp and state public security departments. The information provided by Senasp was accounted for in the period from 2015 to 2022 and the rates were calculated, in the usual way, for each year.

Keywords: intentional homicide; k -means; rate; clusters.

1. Introduction

Measuring criminality is not an easy task due to its breadth and dynamics. Another difficulty is related to the lack of information. In general, information sources have weaknesses such as data availability and coverage, in addition to underreporting. The latter, in general, is linked to the existence of certain occurrences that are more likely to be included in the registration system. The

seriousness of the crime, institutional interests, legal changes in the management of the registry are factors that may affect the estimates provided about criminality. With the outbreak of the COVID-19 pandemic, it is possible to identify sudden variations in some crime indicators between 2020 and 2021, possibly caused by changes in routine and also in the propensity to report crimes. These factors constitute important cautions that must be taken when interpreting criminal trends based on official records.

Official crime statistics are regularly used in many countries to portray public safety. For a crime to be included in official statistics, three successive steps are required: detection of the crime; reporting the crime to law enforcement authorities; the record in the police report or other official registration document. Furthermore, fluctuations in records do not always reflect variations in the criminal phenomenon. These can be caused by notification changes, more or less intense police activities or by legislative or administrative notifications, in addition to other factors. As determined by the United Nations Office on Drugs and Crime (UNODC, 2019), despite occasional periods with substantial declines in indicators, Brazil has historically had high rates of intentional homicide, both in a regional panorama and when compared to other countries in the Americas. In the South, as well as in a global scenario.

In this sense, Feltran *et al.*, 2022 proposes an explanation for the variations in homicide rates in Brazil between the years 2000 and 2019. Based on the comparison of ethnographic experiences lived in the factional universe of four capitals (São Paulo, Porto Alegre, São Luís and Maceió), the authors proposed two analytical strategies: the disaggregation of quantitative series of homicide rates of victim profiles; the construction of historical synopses of local factional conflicts. It was demonstrated how homicide rates, in specific sociodemographic profiles, oscillate from changes in local factional conflicts and pull variations in aggregate rates. Portella *et al.*, 2019 verified the association between intentional homicide, drug trafficking and social indicators in Salvador, capital of Bahia, Brazil, in 2010. This is an ecological study based on registered cases of intentional homicide and drug trafficking by the Civil Police of Bahia with social indicators. The authors used in the regression model considering the negative binomial distribution for comparisons between the coefficient of intentional homicide, drug trafficking and the proportion of black men aged 15 to 49 years in Salvador, Bahia. In the association, it was observed that in neighborhoods with a proportion of black men aged 15 to 49 over 60%, this coefficient increased by 89% ($\geq 60\%$ and $\leq 80\%$) and 87% ($> 80\%$), compared to neighborhoods with proportions less than 60%. As for drug trafficking, there was a statistically significant average increase of 40% in this coefficient in neighborhoods with 5 or more cases, compared to neighborhoods with less than 5 cases.

Concerned with knowing the spatial pattern of the distribution of intentional homicides in João Pessoa, capital of Paraíba, between 2011 and 2016, Oliveira *et al.*, 2019 used the local Moran index with empirical Bayesian smoothing to describe this spatial pattern. Significant spatial autocorrelation was detected. It was identified that the highest risk areas were distributed across the west, northwest, southeast and extreme south zones, mostly in areas of low economic status. Lesser risk areas were located in areas of medium/high socioeconomic status. Loureiro & Silva, 2012 assessed what drives people to commit intentional homicide in Brazil, Colombia, El Salvador, Guatemala and Jamaica. They realized, in a study of panel data, that the determining factors for the decrease in the practice of this crime are developed. These increase in income inequality. Risso, 2014 described some of the actions carried out in São Paulo that culminated in a drop of more than 70% in homicide rates between 2000 and 2010. Between 2012 and 2013, still in São Paulo, the author reveals that one in every five homicides in the same city were committed by a police officer.

Many other justifications for looking with due attention to the problem of the high number of murder victims in Brazilian federative units could be described here. However, we believe that the information presented so far is sufficient. Aiming to present the use of statistical techniques to assist in directing financial aid from the competent bodies in order to reduce the rate per 100,000

inhabitants of this type of crime in the FU's, we decided to use the non-hierarchical grouping technique known as k -means, to supervise data. Thus, the FU's were grouped according to similar rates of intentional homicides and, later, for each of the three groups formed, a detailed description was carried out. It was possible to notice that the three different groups have different characteristics in relation to three variables: per capita gross domestic product, schooling rate, the percentage of social inclusion of the black population and job creation. The FU's with higher rates have the lowest levels for such characteristics.

The statistical technique for multivariate data called cluster analysis Fávero & Belfiore (2019) and Ferreira (2018) with the aim of grouping similar states due to some characteristics so that the similarity of the states that make up each group is maximized and the similarity between the groups is minimized. We chose to use the non-hierarchical cluster analysis known as k -means, having the advantage of allowing the visualization of clusters without overlapping for a very large number of observations and the disadvantage that the choice of the number of groups must be made a priori. However, this choice is "pseudo-arbitrary", since there are several criteria that help in this choice. Here, we use the "elbow" Kodinariya & Makwana, 2013 criterion.

Thus, section 2 presents the methodology used here. A detailed and reasoned description of the k -means method. We will not present the definition of the calculation of the rate per 100,000 inhabitants, the calculation of GDP and other characteristics because we believe that these concepts are widespread in the literature, facilitating a prior consultation by the reader. Then, section 3 presents our relevant results and discussion. Section 4 brings a brief conclusion on the subject exposed here. Finally, references for consultation are presented.

2. Materials and Methods

Intentional lethal violent crimes comprise the criminal records of intentional homicide, attempted murder, robbery followed by death (murder) and bodily injury followed by death. Data are described and analyzed from a historical series whose records are from January 1, 2015 to December 31, 2022. They constitute the crime of intentional homicide (Senasp):

"Morte de alguém em que há indício de crime ou sinal de agressão externa, exceto "feminicídio", "lesão corporal seguida de morte", "roubo seguido de morte (latrocínio)" e crimes culposos; b) Morte violenta provocada por acidente de trânsito, desde que haja dolo; c) Morte com indício de crime ou sinal de agressão externa qualificada como "encontro de ossada", "encontro de cadáver", "morte a esclarecer", "morte suspeita", "morte por causa desconhecida" e congêneres são classificadas como homicídio."

Also, the population of the resulting years analyzed were taken from the "Instituto Brasileiro de Geografia e Estatística - IBGE".

2.1 Methods

In cluster analysis, the objective is to classify observations of a dataset according to their similarities (Everitt *et al.*, 2011), in which similar observations are allocated in the same group and, therefore, those belonging to different groups are considered dissimilar. The similarity between the observations is quantified by means of a proximity metric that can be, for example, the Euclidean quadratic distance between observations x_i and the cluster centroid \bar{x} .

The Euclidean quadratic distance

It is a distance or metric that aims to measure the distance between two points located in a dimensional space using the Pythagoras theorem. According to Fávero & Belfiore (2019), its expression is given by

$$d_{(x_i, \bar{x})}^2 = \sum_{i=1}^p (x_i - \bar{x})^2, \quad (1)$$

where p represents each observation within a k cluster and \bar{x} represents the centroid of the k cluster. This distance is often used when the variable under study has little dispersion.

Another important aspect of the k -means clustering technique is the a priori choice of the number of clusters which, as already mentioned (see section 1), can be defined by the “elbow” criterion.

The elbow criterion

The “elbow” (Bholowalia & Kumar, 2014) criterion aims to define the number of clusters that should be chosen “a priori” so that the total variation within the cluster, known as “WSS” (sum of total squares within the cluster) is minimized and is given by the following expression

$$\text{minimize} \left(\sum_{i=1}^K \text{WSS}(k_i) \right), \quad (2)$$

where k_i is the i th cluster. The total sum of squares within the cluster (“WSS”) measures the compactness of the cluster and we want its value to be as small as possible. Thus, we can use the following algorithm to define the ideal clusters:

1. calculate the value of “WSS” for different values of k , for example, $k = 1, 2, \dots, 10$;
2. plot the “WSS” curve as a function of previously chosen values for k ;
3. the location of an “elbow” in this graph is generally considered to be an indicator of the appropriate number of clusters.

This algorithm was adapted from Kodinariya & Makwana, 2013. The downside is that this criterion has problems when the number of clusters cannot be unequivocally identified. However, this criterion is still the most used. Finally, we introduce the non-hierarchical method k -means.

k -means method

This is a widely used method among partition methods. In this method we must “a priori” inform the number of k clusters and the observations are grouped into these k clusters using an objective function at the criterion. Based on Fávero & Belfiore (2019), Johnson, Wichern, *et al.*, 2002, we present a logical sequence of steps.

1. Initially define the number of clusters as well as their respective centroids. Thus, the main objective is to divide the dataset observations into k clusters, $k = 1, 2, 3, \dots$, where the observations within each cluster are closer to each other when compared to any other observation that belongs to a different cluster. The observations were arbitrarily allocated in the k clusters so that their respective centroids can be calculated;
2. The next step is to check if a given one is closer to another centroid using the Euclidean distance between points, as in 1 and, if so, reallocate it in this cluster. At this moment, a cluster has just lost this observation and, therefore, the centroids of the cluster that receives it and the cluster that loses it must be recalculated;
3. Step 2 must be repeated until it is no longer possible to relocate any observation due to its proximity to a centroid of another cluster.

Figure 1 illustrates these logical steps. Figures 1(a), 1(b) illustrate steps 1. Figures 1(c) and 1(d) illustrate step 2. Step 3 is illustrated in Figures 1(e).

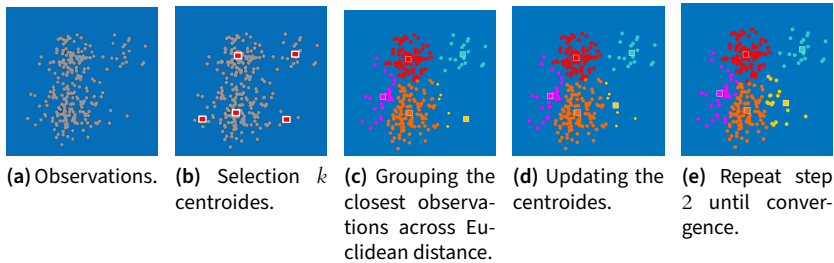


Figure 1. Algorithm of the k -means method. The dots represent the observations. The squares represent the centroids of the clusters.

Each centroid \bar{x} must be recalculated whenever an observation p is added or deleted in a new centroid in the respective cluster, based on the following expressions

$$\bar{x}_{\text{new}} = \frac{N\bar{x} + x_p}{N + 1}, \text{ se a observação } p \text{ é incluída no cluster} \quad (3)$$

$$\bar{x}_{\text{new}} = \frac{N\bar{x} - x_p}{N - 1}, \text{ if observation } p \text{ is excluded from the cluster} \quad (4)$$

where N refers to the number of observations in the cluster, \bar{x} refers to the relocation centroid of this observation, and x_p refers to the observation p to be included in a new cluster. The “a priori” choice of the number of clusters k is necessary to avoid a large computational effort, because when we consider all possible clusters we can obtain a very large number of possibilities.

Once this procedure has been carried out, it is important to define whether the clusters found make sense, that is, whether the variability within clusters is really smaller than the variability between clusters. This fact is equivalent to testing the following hypotheses

$$\begin{aligned} H_0 &: \text{the variable has the same average in all formed groups} \\ H_1 &: \text{the variable has different means in each of the formed groups.} \end{aligned} \quad (5)$$

However, one can use the Elbow criterion (see section 2.1) as a criterion for deciding on these hypotheses. Some authors mention a test statistic that under the null hypothesis, as in 5, follows a distribution F with $k - 1$ degrees of freedom in the numerator and $n - k$ degrees of freedom in the denominator, being n the number of observations. This method would be used in analysis of variance and therefore has the disadvantage of requiring multivariate normality of the data, which is not always guaranteed (Kodinariya & Makwana, 2013). Our main results are presented in the next section.

3. Results and Discussion

Our main results are all based on calculating the rate per 100,000 inhabitants. So, Figure 2 presents the distribution of these rates, for each FU, over the period studied. Bars with colors of the same size indicate similar rates and, therefore, indicate possible groupings of FU's by similar rates. We highlight that in São Paulo, in 2015, there were no records of information by Sinesp. The registrations started only in December 2016. This fact stems from the beginning of the implementation of the Sinesp system in Brazil.

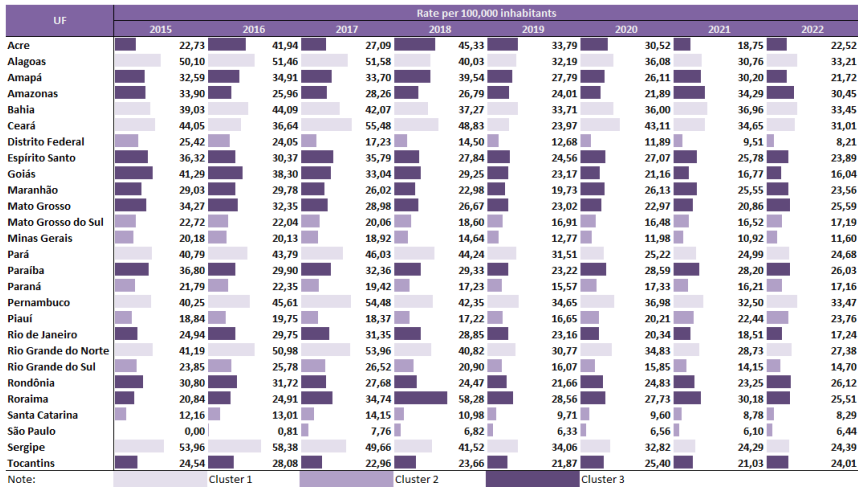


Figure 2. The distribution of FU's rates over the period 2015-2022. Bars with colors of the same size indicate the FU's with similar rates per 100,000 inhabitants.

The information contained in Figure 2 is sufficient for us to visualize three clusters in the Brazilian FU's, by the similarity between the rates per 100,000 inhabitants in the current period. However, there is a criterion for the optimal choice of the maximum number of clusters that is widespread in the literature: the elbow criterion (see section 2.1). Figure 3 suggests three clusters, an expected result (see Figure 2).

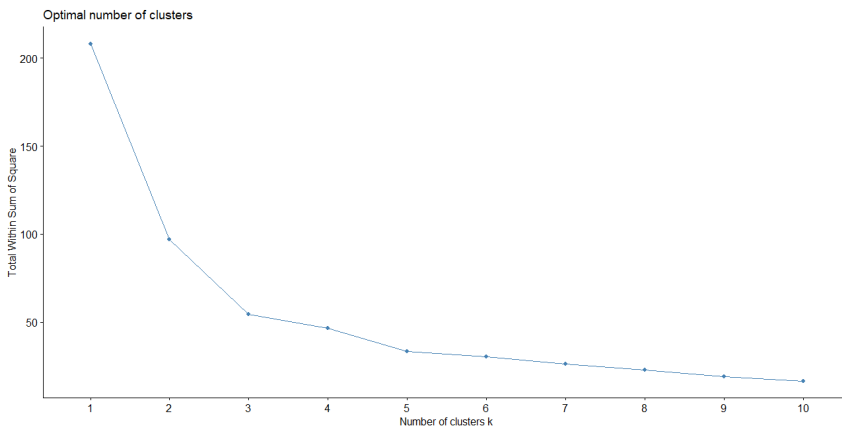


Figure 3. Choice of the optimal number of clusters using the elbow criterion.

Table 1 presents a detailed description of each of the three clusters found. For cluster 1, we have 8 FU's, 3 belonging to the southern region, 2 belonging to each of the central-west and southeast regions and only 1 from the northeast region. Cluster 2 is formed by 12 FU's, 6 FU's belonging to the North region, 2 belonging to each of the respective Midwest, Southeast and Northeast regions. Cluster 3 has 7 FU's in its formation, 6 belonging to the northeast region and only one belonging to the northern region.

Figures 4 show that only the FU's that form clusters 1 have homicide rates below 16 victims per 100,000 inhabitants or at the limit throughout 2015-2022.

With this information in hand, we present our conclusion regarding this study.

Table 1. Description of clusters found.

Cluster	Lenght	FU	Mean (2015-2022)			T.A. (km ²)	D.D. (hab/km ²)	E.E.S.	H.D.I. (2021)	H. I. (per capita)	T. V. (2022)
			Population	Victims	Rate						
1	8	DF	2,991,134	456.37	15.26	5,760.784	489.01	369,128	0.814	2,913	2,021,627
		MS	2,762,473	518.25	18.76	357,142.082	7.72	391,975	0.742	1,839	1,824,708
		MG	21,097,249	3,184.38	15.09	586,513.983	35.02	2,407,107	0.774	1,529	13,028,883
		PR	11,387,509	2,089.38	18.35	199,298.981	57.42	1,348,296	0.769	1,846	8,575,905
		PI	3,267,470	642.50	19.66	251,755.481	12.99	459,871	0.69	1,11	1,385,426
		RS	11,348,784	2,234.13	19.69	281,707.151	38.62	1,257,992	0.771	2,087	7,869,630
		SC	7,116,856	597.17	8.39	95,730.69	79.49	90,024	0.792	2,018	5,974,106
		SP	45,707,511	2,348.00	5.14	248,219.485	178.96	5,396,803	0.806	2,148	32,293,191
2	12	AC	875,494	264.63	30.23	164,173.429	5.06	153,015	0.710	1,038	334,377
		AP	837,098	256.63	30.66	142,470.762	5.15	133,839	0.688	1,177	232,691
		AM	4,111,259	1,158.88	28.19	1,559,255.881	2.53	702,763	0.700	965	1,069,794
		ES	3,994,133	1,152.38	28.85	46,074.448	83.20	503,003	0.771	1,723	2,248,960
		GO	6,968,274	1,888.13	27.10	340,242.859	20.74	855,021	0.737	1,619	4,542,235
		MA	7,054,329	1,786.63	25.33	329,651.496	20.55	1,112,636	0.676	814	2,031,236
		MT	3,462,204	925.50	26.73	903,208.361	4.05	486,568	0.736	1,674	2,568,240
		PB	4,006,799	1,172.75	29.27	56,467.242	70.39	540,919	0.698	1,096	1,523,167
		RJ	17,206,356	4,166.00	24.21	43,750.425	366.96	1,945,408	0.762	1,971	7,475,503
		RO	1,766,43	463.88	262.61	237,754.172	6.65	244,815	0.700	1,365	1,154,287
		RR	590,334	185.00	31.34	223,644.53	2.85	103,123	0.699	1,242	263,345
TO	1,568,896	374.13	23.85	277,423.627	5.45	227,743	0.731	1,379	830,733		
3	7	AL	3,329,242	1,351.75	40.60	27,830.661	112.38	458,782	0.684	935	1,034,187
		BA	14,838,257	5,608.63	37.80	564,760.429	25.03	1,946,957	0.691	1,010	4,887,673
		CE	9,102,314	3,609.63	39.66	148,894.447	59.05	1,161,434	0.734	1,050	3,625,994
		PA	8,555,970	2,992.38	34.97	1,245,870.704	6.51	1,389,983	0.690	1,061	2,478,988
		PE	9,526,046	3,807.25	39.97	98,067.877	92.37	1,249,850	0.719	1,010	3,439,164
		RN	3,491,904	1,342.50	38.45	52,809.599	62.53	447,692	0.728	1,267	1,495,094
		SE	2,287,978	907	39.64	21,938.188	100.72	320,638	0.702	1,187	907,388

Note: T.A. - Territorial area; D.D. - Demographic Density; E.E.S. - Enrollments in elementary school; H.D.I. - Human development index; H.I. - Household monthly income per capita; T.V. - Total vehicles



(a) Rate per 100,000 inhabitants of the FU's forming cluster 1. (b) Rate per 100,000 inhabitants of the FU's forming cluster 2. (c) Rate per 100,000 inhabitants of the FU's forming cluster 3.

Figure 4. Rate per 100,000 inhabitants of the FU's forming each cluster. The gray dashed line indicates the limit imposed by the 2022 National Public Security Plan - 16 victims per 100,000 inhabitants.

4. Conclusions

The clustering technique known as *k*-means proved to be a useful tool to discriminate and group possible clusters in Brazilian FU's through similarities of intentional homicide rates. The lack of availability of socioeconomic data was a determining factor in the scarcity of information in the description of the clusters. So, we use, as far as possible, that information that we think is relevant. It is possible to notice that the FU's forming cluster 1 have better socioeconomic characteristics than those forming clusters 2 and 3. As future studies, the same idea applied to information from municipalities can be extended. We did not do so because we are waiting for Senasp to update this information and there is no prediction of when it will occur.

Acknowledgments

We would like to thank the Institute for Applied Economic Research – IPEA and the Directorate of Studies and Policies of the State, Institutions and Democracy (Diest) for their substantial support in the construction of this paper. Likewise, we would like to thank the National Secretariat for

Public Security (Senasp) and the Ministry of Justice (MJ) for openly providing the data used here. We are grateful to the reviewers and editors for their comments and suggestions that help to improve the quality of the paper.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The authors must include the CRediT- Contributor Role Taxonomy considering each author contribution for the paper, using the following roles (for more information, access <https://credit.niso.org/>):

Conceptualization: ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Data curation:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Formal analysis:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Funding acquisition:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Investigation:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Methodology:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Project administration:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Software:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Resources:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Supervision:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Validation:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Visualization:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Writing - original draft:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O. **Writing - review and editing:** ALVES, H.J.P.; FERNANDES, F. A.; SILVA, E.M.; BATISTA, B. D.O.

References

1. Bholowalia, P. & Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* **105** (2014).
2. Everitt, B. S., Landau, S., Leese, M. & Stahl, D. *Cluster analysis 5th ed* 2011.
3. Fávero, L. P. & Belfiore, P. *Data science for business and decision making* (Academic Press, Cambridge, MA, USA, 2019).
4. Feltran, G., Lero, C., Cipriani, M., Maldonado, J., Rodrigues, F. d. J., Silva, L. E. L. & Farias, N. Variations in homicide rates in Brazil: an explanation centred on criminal group conflicts. *Dilemas: Revista de Estudos de Conflito e Controle Social* **15**, 311–348 (2022).
5. Ferreira, D. F. *Estatística multivariada* 3rd ed., 624 (Editora UFPA, Universidade Federal de Lavras, 2018).
6. Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis* **8** (Prentice hall Upper Saddle River, NJ, Upper Saddle River, 2002).
7. Kodinariya, T. M. & Makwana, P. R. Review on determining number of Cluster in K-Means Clustering. *International Journal* **1**, 90–95 (2013).
8. Loureiro, P. R. A. & Silva, E. C. What causes intentional homicide? *Journal of International Development* **24**, 287–303 (2012).
9. Oliveira, A. L. S. d., Luna, C. F., Quinino, L. R. d. M., Magalhães, M. d. A. F. M. & Santana, V. C. d. M. Spatial analysis of intentional homicides in João Pessoa, Paraíba, Brazil, 2011–2016. *Epidemiologia e Serviços de Saúde* **28**, e2018184 (2019).

10. Portella, D. D. A., Araújo, E. M. d., Oliveira, N. F. d., Chaves, J. M., Rocha, W. d. J. S. d. F. & Oliveira, D. D. Intentional homicide, drug trafficking and social indicators in Salvador, Bahia, Brazil. *Ciencia & saude coletiva* **24**, 631–639 (2019).
11. Risso, M. Intentional homicides in São Paulo city: A new perspective. *Stability: International Journal of Security and Development* **3** (2014).