computation

MDPI

Article

Dendrograms for Clustering in Multivariate Analysis: Applications for COVID-19 Vaccination Infodemic Data in Brazil

Maria da Penha Harb [1,*], Lena Silva [2], Thalita Ayass [1], Nandamudi Vijaykumar [3], Marcelino Silva [4] and Carlos Renato Francês [1]

[1] Institute of Technology, Federal University of Para, Belem 66075-110, Brazil
[2] Center for Exact Sciences and Technology, University of Amazon, Belem 66060-902, Brazil
[3] National Institute for Space Research, São José dos Campos 12227-010, Brazil
[4] Institute of Engineering and Geosciences, Federal University of West Para, Santarem 68040-255, Brazil
* Correspondence: mpenha@ufpa.br

Abstract: Since December 2019, with the discovery of a new coronavirus, humanity has been exposed to a large amount of information from different media. Information is not always true and official. Known as an infodemic, false information can increase the negative effects of the pandemic by impairing data readability and disease control. The paper aims to find similar patterns of behavior of the Brazilian population during 2021 in two analyses: with vaccination data of all age groups and using the age group of 64 years or more, representing 13% of the population, using the multivariate analysis technique. Infodemic vaccination information and pandemic numbers were also used. Dendrograms were used as a cluster visualization technique. The result of the generated clusters was verified by two algorithms: the cophenetic correlation coefficient, which obtained satisfactory results above 0.7, and the elbow method, which corroborated the number of clusters found. In the result of the analysis with all age groups, more homogeneous divisions were perceived among Brazilian states, while the second analysis resulted in more heterogeneous clusters, recalling that at the start of vaccinations they could have had fear, doubts, and significant belief in the infodemic.

Keywords: dendrogram; infodemic; COVID-19; Google Trends; multivariate analysis

Citation: Harb, M.d.P.; Silva, L.; Ayass, T.; Vijaykumar, N.; Silva, M.; Francês, C.R. Dendrograms for Clustering in Multivariate Analysis: Applications for COVID-19 Vaccination Infodemic Data in Brazil. *Computation* **2022**, *10*, 166. https://doi.org/10.3390/computation10090166

Academic Editors: Simone Brogi and Vincenzo Calderone

Received: 8 August 2022
Accepted: 5 September 2022
Published: 19 September 2022

1. Introduction

Humanity has been seriously affected by the COVID-19 pandemic that originated in late 2019 in Wuhan, Hubei province in China, caused by the SARS-CoV-2 virus [1,2]. It was declared a pandemic on 11 March 2020, by the World Health Organization (WHO) [3]. According to the WHO data [4], by May 2022, more than 533 million cases of the disease had already been confirmed, causing more than 6 million deaths worldwide.

Both the impact of an entirely new and unknown disease, which was already immense, and the lack of information associated with it allowed false and dubious information to quickly appear and spread across various social media platforms, newspapers, and magazines [5]. Thus, the COVID-19 pandemic was accompanied by a massive and widespread wave of disinformation about the disease that can be described as an infodemic [6].

The WHO explains that infodemics are an excessive amount of information about a problem, making it extremely hard to identify a solution. They can spread both misinformation and disinformation (accidentally and deliberately false information, respectively) as well as rumors during a health emergency. Infodemics can severely affect or damage an effective public health response and create confusion and mistrust among people [7].

The battle against the COVID-19 pandemic and infodemic continues. A long-awaited step was the development of effective vaccines, which was highly anticipated, and several vaccines are now available. The development and availability of vaccines are not the only

obstacles to overcome from a public health perspective. The increasing acceptance of the vaccine by the population is also paramount to designing public health measures and reaching a considerable proportion of vaccinated population to achieve herd immunity [8].

However, there is a process of rejection or delay in accepting vaccines, which can be affected by the variables of trust, compliance, and convenience, and directly influences the historical context of vaccination [9]. Such resistance is known as vaccine hesitancy, and before the pandemic, in 2019, the WHO identified it as one of the major threats to global health [10]. In addition to fake news, other factors impact the drop in vaccine coverage, such as social, cultural, religious, and economic issues in which the population is involved, and this can influence whether the population goes to vaccination centers [11].

Following the evolution of the pandemic, until May 2022, Brazil was the third-most affected country in the total number of cases and second in the total number of deaths [12], ranked 34th in vaccination (counting both first and second vaccine doses) [13]. Brazil applied its first vaccine against COVID-19 on 17 January 2021. In three months, 5.32% of the population had received one of the two necessary doses and only 2% were fully immunized [13]. The pace was very slow compared to countries like Israel, the United Kingdom, and Chile [14]. This was due to problems and delays in vaccine purchases (and when a campaign started, doses ran out), doubts about the effectiveness of the results, and false and dubious information circulated on the internet and social networks and in speeches by the President of Brazil, who was considered a denialist and anti-vaccine [14,15].

Because of the above, the purpose of this work is to carry out a spatial analysis of COVID-19 vaccination by the Federation Units in Brazil from the contribution of the cluster analysis technique, applying multivariate techniques using indicators that cover the infodemic data of vaccination from COVID-19 in Google Trends (GT) searches, internet proliferation in states, and data on deaths and the number of COVID-19 cases.

Cluster analysis makes it possible to group cases or variables into a homogeneous group based on their similarity. Each object is like the others in the group, maximizing homogeneity within the group and maximizing heterogeneity between groups. Dendrograms are used to cluster the states and, thus, generate new divisions of the regions, different from the geographic regions that already exist in Brazil: first carried out for an analysis of all age groups available and second with an analysis of the vaccination of the elderly 64 years or older.

Brazil is the fifth largest country in the world geographically and the sixth in population; it has 5570 municipalities divided into 27 federative units (26 states and one Federal District (DF)), which are grouped into five geographic macro-regions (Central-West, Northeast, North, Southeast, and South). After the results obtained from the techniques used, the new division of the groupings of the states will be visualized through the behavior of the population.

2. Related Work

The internet is revolutionizing the way epidemic intelligence is collected and offers solutions to some of these challenges. Freely available sources of information can allow us to detect disease outbreaks earlier with reduced cost and greater transparency in reporting [16]. Search engines have become pervasive in recent years, retrieving information easily on a variety of topics, ranging from customer service to general information. In addition to these research interests, there is a growing interest in obtaining health advice or information. In this respect, health policy authorities have begun to identify internet search engines as potential indicators for surveillance and health, such as the GT, a repository of publicly available information on user research patterns and real-time data [17].

In Mangono et al. [18], GT was used to provide insights and potential indicators of important changes in information-seeking patterns during COVID-19 with various pandemic-related terms, such as: coronavirus symptoms, urgent care near me, health insurance, social distancing, and "Chinese virus", among others. They compared the surveys with new monthly Medicaid orders (an application that provides health coverage

to Americans), and used principal component analysis to identify research patterns in the GT.

Rovetta and Bhagavathula [19] investigated online search behavior related to the COVID-19 outbreak and the "infodemic nicknames" circulating in Italy using GT. The titles of articles from the most read national newspapers and government websites were surveyed to investigate the extent and attitudes of several related infodemic nicknames for COVID-19. They defined "infodemic nicknames" as substantially erroneous information, which gave rise to misinterpretations, fake news, episodes of racism, or any other form of misleading information that circulated on the internet. They concluded that Google search query data reflects growing regional and population interest in the pandemic. Searches related to disinfectants, face masks, health newsletters, vaccines, and symptoms related to COVID-19 were the main search keywords. However, many infodemic nicknames circulated in Italy. They also conclude that internet research interest in COVID-19, both at the regional and city levels in Italy, was influenced by tradition, electronic newspapers, and printed media coverage.

In Ceron et al. [20], the authors explored news-checking initiatives in Latin America, using a Markov-based computational method to group tweets into topics and identify their diffusion among different datasets about false information related to COVID-19 across regions, comparing if there was a pattern for Argentina, Brazil, Chile, Colombia, Mexico, and Venezuela.

Multivariate statistical methods were used by Custodio et al. [21], where they perceived that health measures led to a significant reduction in air pollution, but on the other hand, the impact of these measures in aquatic environments was poorly analyzed. In this context, multivariate statistical methods were employed to evaluate the water quality of the rivers of the Mantaro River basin and heavy metal contamination indices during the health crisis associated with the COVID-19 pandemic. The techniques employed were principal component analysis and hierarchical cluster analysis according to Spearman's correlation, which generated a dendrogram where the five chemical elements were grouped into two statistically significant groups, observing a significant increase in the critical value of contamination.

Computational and statistical techniques were used to analyze the heterogeneous spread of the pandemic and estimate the death rates from COVID-19 [22–24]. In Silva et al. [22], the authors estimated the effective reproduction rate number for each epidemiological week in Brazil and designed scenarios based on these values, concluding that the only way to flatten the curve is to decrease the reproduction rate. The work of Shafiq et al. [23] was applied to data from Italy and the results revealed that the model of artificial neural networks is an excellent engineering tool to predict survival and mortality rates, presenting more satisfactory and better results than other studies found in the literature.

Multivariate analysis techniques and dendrograms on the pandemic, as well as data from Brazil, are being used for grouping behaviors. James et al. [25] compared and contrasted data from the USA, India, and Brazil, looking at the trajectories of cases, deaths, and mortality rates, which were analyzed state by state. Dendrograms were used and revealed a similar cluster structure between the USA and India. Both countries had a dense majority cluster and a small collection of outliers. Brazil, on the other hand, presented a quite different structure, with two similarly sized clusters that contained most of the elements and then some outliers.

James and Menzies [26] presented a mathematical framework for determining the behavior of the second outbreak of COVID-19 cases in the United States based on a collection of time series. The data were grouped (dendrograms), identifying the different outbreak behaviors, and in the appendix of the work, the authors applied the technique to the data from Brazil and concluded that the second outbreaks were moderate and of comparable severity to the first outbreaks. This is similar to the USA, which also noted significant similarities based on geography.

In the research by Harb et al. [27], a multivariate analysis was performed on a database of COVID-19 infodemic terms in Brazil over 18 months (1 January 2020 to 30 June 2021), including socioeconomic and political variables. The infodemic terms were divided into five groups, and the analysis was performed every 3 months during the evolution of the pandemic. The study concludes that denial about the pandemic and the influence of political leadership can influence the search for infodemic information, contributing to disorganization in the control of the disease and prevention measures.

In this work, using the methodology of Harb et al. [27], we carried out an approach by combining the database of infodemic vaccination terms, extracted from research in the GT, related to variables of vaccination numbers (first and second doses), in the entire year of 2021 for Brazil. Multivariate analysis with dendrograms and the elbow method was performed to group the states into similar patterns of behavior in the two analyses performed.

3. Methodology

The methodology applied in this work can be seen in the flowchart in Figure 1, based on Harb et al. [27] and Braz et al. [28], and the data obtained were obtained from the 27 federative units (26 states and Distrito Federal (DF)) for the year 2021. Python language was used to perform both data manipulation and data clustering steps, and Tableau software (version 2021.3.3, Tableau Software, Mark Nelson, Seattle, WA, USA) was used to reproduce the map of Brazil with the generated clusters.

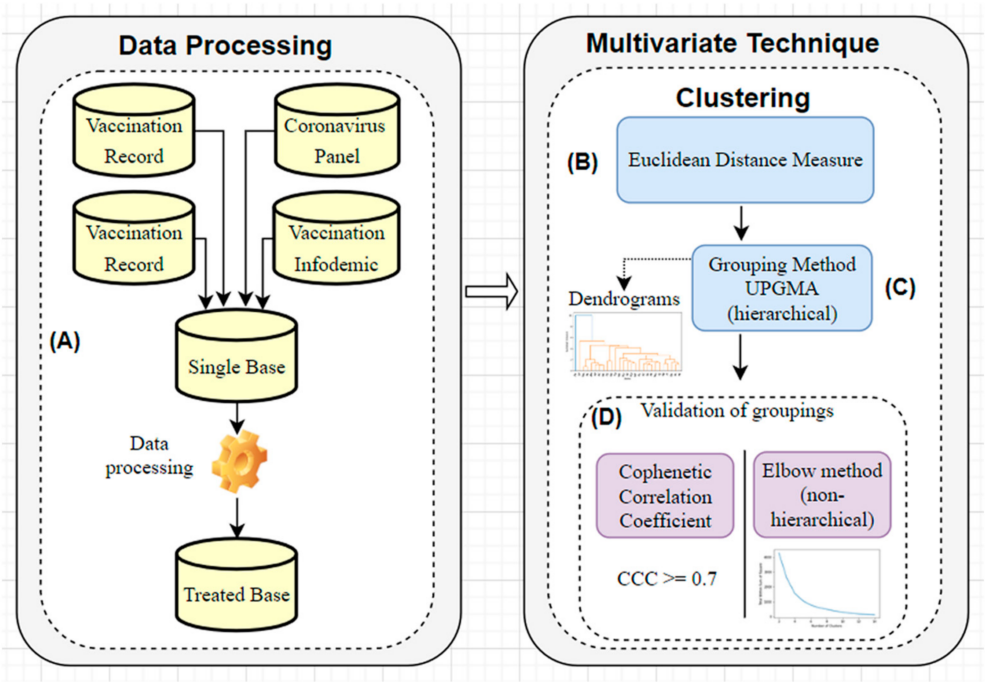

Figure 1. Flowchart of the methodology of the multivariate technique adopted. Processing the data: (**A**) selection and treatment of variables. Multivariate Technique: (**B**) selection of the distance measure; (**C**) selection of the clustering algorithm and (**D**) validation of the clusters.

The two main steps are presented in Figure 1, processing the selected data and clustering. The first step (A) is the extraction and treatment of the variables, generating a single dataset. The subsequent step applies and tests the best parameters for clustering (B) and (C), and at the end, performs cluster validation with two techniques (D).

3.1. Selection and Treatment of Variables for the Database (A)

The first step of the flowchart is the creation of the database, with the selection of variables to be used, the identification of outliers, and data standardization. Four databases available on the internet were used. Three are public databases from the government, and

the last database is data extracted from GT. The following steps were considered: data collection, data processing, data mining, data interpretation, and validation. The databases searched were:

1. COVID-19 vaccination records [29]

For each state, three files were made available. After collecting and joining the files, the vaccination fields were selected by age group, counting the people who completed the vaccination cycle (1st and 2nd doses). To achieve a better relationship among the states, the rate per 100,000 inhabitants was calculated for these fields: data divided by the population of each state, multiplied by 100,000.

2. Coronavirus Panel [30]

The fields with information on the number of COVID-19 deaths and number of new cases were selected and the annual average per state was performed.

3. Internet access density [31]

Information on fixed broadband and mobile telephony was selected and the average was determined per each state.

4. Vaccination Infodemic [32]

The infodemic terms about vaccination [27–31,33–36] were selected. A search was performed on GT by the state for each term. The research was carried out using the following filters: geographic, which was used for each Brazilian state; period, which was the chosen year 2021; and categories, which were chosen for all categories and research groups using web searches. The values returned are from 100 (represents a term's peak popularity) to 0 (means there was not enough data about the term). The search was carried out on 1 May 2022, and for each term, the results found in related subjects and related searches were evaluated in order to verify if the term was related to non-infodemic content. As a result, some terms were dropped out.

The infodemic terms selected were: jacare vaccine, turning into jacare, doria vaccine, DNA vaccine, mutated DNA vaccine, vaccine kills, vaccine kills COVID, COVID cancer, COVID cancer vaccine, alcoholic drink COVID vaccine, liquid chip, chip vaccine, COVID hiv vaccine, COVID vaccine Alzheimer, magnetic COVID vaccine, COVID vaccine CoronaVac, COVID fetus vaccine, magnetic COVID vaccine, aluminum coronaVac, vaccine causes autism, vaccine causes impotence, and CoronaVac squint. Some terms are very specific in Brazilian news, such as the term "turning into alligator", researched after the President of Brazil said that after taking the vaccine, a person would become an alligator [37].

After the process of selecting the variables in the databases, outliers were identified, as the clustering technique is sensitive to outliers [38]. However, after analyzing the database, it was decided these values should not be removed. The reason for this is that outliers may form isolated clusters, which, in the case of Brazilian states with more striking characteristics, can happen.

With the database already formed, the standardization of variables was carried out because the use of measurement scales in different magnitudes can distort the analysis, and the most chosen form, among so many techniques, was the standardization z-score, with a mean of zero standard deviation 1 [39,40], shown in Equation (1),

$$Z = \frac{x - \mu}{SD} \quad (1)$$

where x is a data value, μ is the average of the values of the interval, and SD is the standard deviation.

The final database contains information from the 27 federative units for the year 2021. Twelve attributes were selected and treated, as shown in Table 1.

Table 1. Description of the variables selected in the database.

Name Variable	Type	Value
state	text	Brazilian State
number12_17	number	Age range 12 to 17 [1]
number18_64	number	Age range 18 to 64 [1]
number65_69	number	Age range 65 to 69 [1]
number70_74	number	Age range 70 to 74 [1]
number75_79	number	Age range 75 to 79 [1]
number80_	number	Age over 80
medCasosNovos	number	New COVID-19 cases, on average, per day
medObitosNovos	number	COVID-19 deaths, on average, per day
densblf	number	Density of fixed broadband [2]
denstm	number	Density of mobile phone [3]
infodemia	number	Relative volume of research for infodemic terms

[1] Number of people vaccinated with both first and second doses; [2] Number of accesses divided by the number of households; [3] Number of accesses divided by population.

3.2. Selection of the Measure of Distance or Similarity between Each Pair of Objects (B)

After the database was complete, the step was to select the measure of distance or similarity between each pair of objects. According to Metz [41], this approach builds the clusters so that examples belonging to the same cluster have a high similarity and examples belonging to different clusters have a low similarity. The measure chosen was Euclidean distance, the smallest distance between two components. The smaller the distance, the more similar the observations [42], and it is the most used distance metric in cluster analysis [41].

Equation (2) represents the Euclidean distance, where the distance between two observations (i and j) corresponds to the square root of the sum of the squares of the differences between the pairs of observations (i and j) for all p variables:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} \tag{2}$$

where x_{ik} is the value of the variable k referring to observation i and x_{jk} represents variable k for observation j.

3.3. Selection of the Clustering Algorithm: Hierarchical Method (C)

This step selects the algorithm to maximize the differences between the groups in comparison with the variation within them. They are divided into hierarchical and non-hierarchical methods. According to Berkhin [43], some characteristics must be considered when choosing the clustering algorithm, such as: input types (dissimilarity matrix), attribute types, ability to find groups with different shapes, and outlier tolerance, among others. These characteristics were evaluated, and the hierarchical average method or unweighted pair group method with arithmetic mean (UPGMA) was chosen, which presented the best results in all executions. This method is less sensitive to noise and outliers and is defined by Equation (3):

$$d(i,j) = \frac{1}{|i||j|} \sum_{\substack{x_a \in i \\ x_b \in j}} d(x_a, x_b), \tag{3}$$

where the distance $d(i,j)$ between two groups is given by the average distance between objects of different groups.

One of the main advantages of hierarchical clustering algorithms comes from the representation of their results through dendrograms. A dendrogram is a graphical representation in tree format that presents the hierarchy of the clusters obtained [44]. In our paper, we elaborated on the dendrograms by executing the clustering of the database.

3.4. Validation of the Clusters (D)

At this stage, the quality of the generated clusters was verified through two algorithms. The first is the cophenetic correlation coefficient (*CCC*), which measures the degree of fit between the similarity matrix (phenetic matrix *F*) and the matrix resulting from the simplification provided by the clustering method (cophenetic matrix *C*) [45]. According to the proposal of Rohlf [45], cophenetic correlations >0.7 are admissible for good clusters and it is obtained by Equation (4):

$$CCC = \frac{C\hat{o}v(F,C)}{\sqrt{V(F)V(C)}} \quad (4)$$

The second algorithm is the elbow method, which interprets and validates coherence within cluster analysis, designed to help find the appropriate number of clusters within a dataset (non-hierarchical method). This method allows for evaluation on how the homogeneity or heterogeneity within the clusters varies for the value of each cluster. We can see this "elbow" when plotting its results on a graph, as there is a break in the direction of the curve, possibly informing the number of clusters to be defined [46]. Equation (5) shows the objective function, a squared error function:

$$W(S,C) = \sum_{k=1}^{k} \sum_{X_i \in S_k} |Y_i - C_k|^2 \quad (5)$$

where S is a k-cluster partition of the entity set represented by the vectors Y_i ($i \in I$) in the M-dimensional feature space, consisting of non-empty non-overlapping clusters S_k, each with a centroid C_k (k = 1,2, ... K).

4. Computational Results

The database presented for the cluster analysis consisted of 27 observations and 12 indicators (as shown in Table 1). The data were organized in a spreadsheet and standardized as the indicators have different measurement units or scales and can change or alter the grouping structure.

4.1. Analysis with All Age Ranges of Vaccination

The dissimilarity measure used was the Euclidean distance, and for the composition of the clusters, the Average method was used, which presented the best *CCC* result, as can be seen in Table 2. It shows the comparison of the results of the execution of the most used methods for clustering. For each method, the *CCC* value, number of clusters, and the cut-off value in the dendrograms are presented.

Table 2. Comparison of the results of the execution of the most used methods for clustering.

Methods	*CCC*	Cluster Number	Cut Dendrogram
Average	0.887	6	3.1
Centroid	0.884	6	2.5
Complete	0.734	7	3.6
Single	0.808	7	1.6
Ward	0.647	6	5.0

The cut made on the axis of dissimilarity of the dendrogram was at a height of 3.1, which demonstrated the composition of six probable groups, as can be seen in Figure 2. Maranhão (MA) and Pará (PA) are highlighted as having particularities different from the others and greater similarities to each other in terms of the behavior of the indicators studied for the observed period. Observing the generated dendrogram, the first DF and second São Paulo (SP) groups were considered groups with an isolated state.

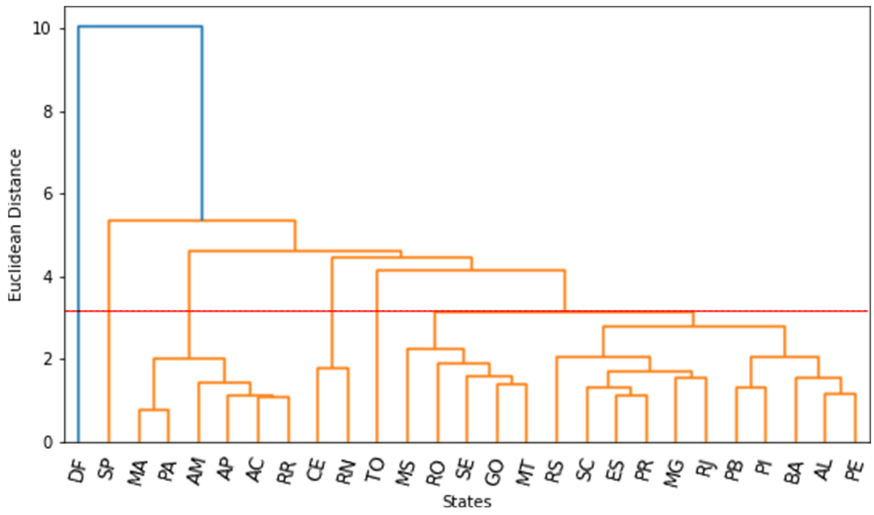

Figure 2. Dendrogram generated for the database with all age groups. The multivariate analysis technique was applied to the 2021 data, grouping Brazil in a different format from the Brazilian regions, which can be observed in the branches of the tree. It presents satisfactory *CCC* (0.887), and six clusters were suggested.

DF is the state with the highest average number of new cases per day of COVID-19 and SP is the largest state in terms of the Brazilian population, and, therefore, they may not have similarities with the other states in the clusters, both with just one state. The third group has six states (MA, PA, Amazonas (AM), Amapá (AP), Acre (AC), and Roraima (RR)), with a representation of 22.22%. The fourth group is composed of two states (Ceará (CE) and Rio Grande do Norte (RN)). The fifth group (Tocantins (TO)), composed of only one state, did not show similarities in behavior with other observations. Finally, the sixth group, the largest in the number of states with sixteen, encompasses states from all Brazilian geographic regions, representing 59.26% of the total states.

To evaluate the generated dendrogram, the *CCC* was employed, presenting a result of 0.887, a value admissible as a good cluster [45]. Thus, the *CCC* result obtained in this research shows an adequate adjustment of the applied clustering method.

The application of the elbow method contributed to defining the value of the number of clusters to be used in the non-hierarchical k-means technique for forming clusters, given that it collaborates in the optimization of clusters [47]. Figure 3 presents the result of the elbow method for the studied database.

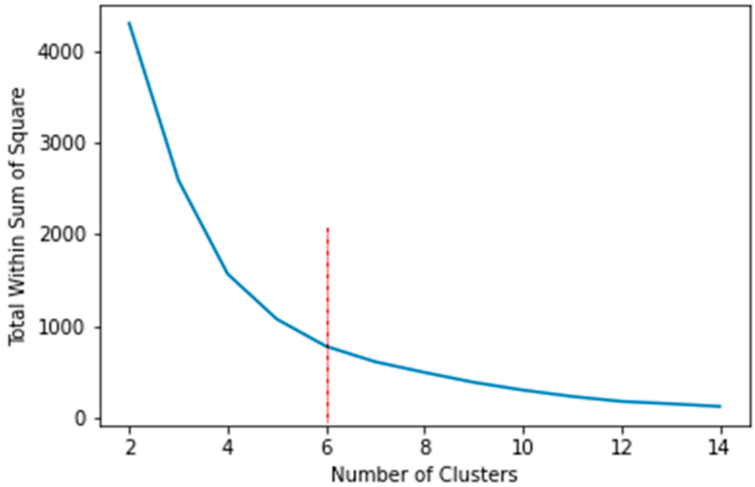

Figure 3. Curve of elbow method generated for the database with all age groups. The number of the cluster was determined by looking at the point position on the "elbow" arm.

The line traces the variation explained as a function of the number of clusters and chooses the elbow of the curve as the number of six clusters to be used.

For better visualization of the dendrogram result, the spatial distribution of the states was performed on the Brazilian map. The map (Figure 4) shows that this technique allowed for the clustering of the states, presenting some dispersed points, possibly because of some local peculiarities distributed within the group. However, there is a predominance of clusters within the same Brazilian region, these being within the same group, according to the method used.

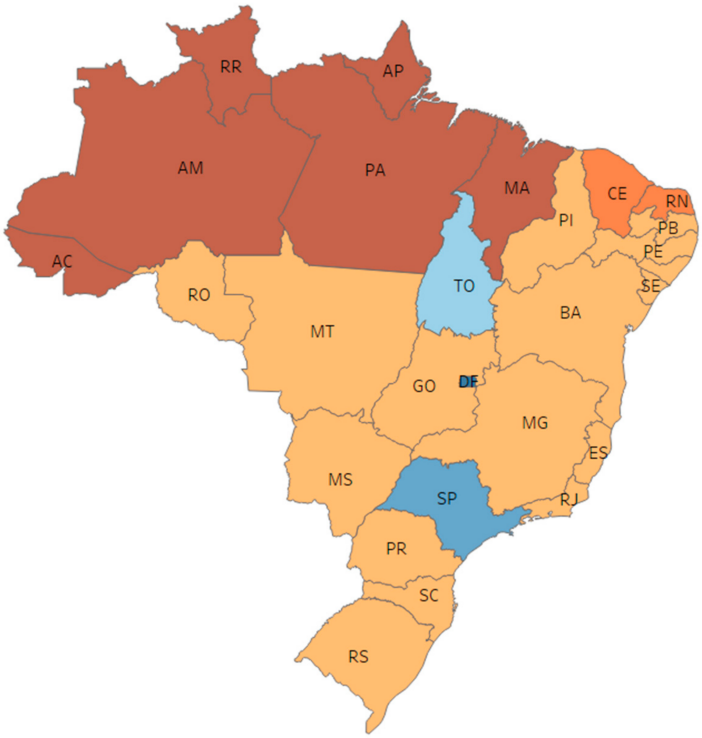

Figure 4. Map of Brazil generated from the dendrogram result with six clusters (with all age groups). Two large clusters contain 80% of the Brazilian states, presenting a great homogeneity in the behavior of the population.

It was noticed that two of the six groups generated present the majority of Brazilian states, indicating similarities in the behavior of the indicators studied. The group with the states MA, PA, AM, AP, AC, and RR are the states with less internet proliferation [48]. In the largest group, with sixteen states, a possible advance in vaccination is observed in Brazil for the year 2021, especially in the South and Southeast regions, which have a high percentage of the immunized population, but areas in the North and Northeast regions still have low immunization rates for COVID-19 [49].

The North and Northeast are places with a lower Human Development Index, younger populations, less educated residents, lower income, and more residents of small towns. In these places, the end of the pandemic seemed farther away than it did for large São Paulo centers (individual clusters), which already have high vaccine coverage with two doses, according to scientists [49], and also showed a lower rate of infodemic searches.

The state of DF, on the other hand, had the highest number of infodemic surveys on vaccination in 2021.

4.2. Analysis with Vaccination of the Elderly 64 Years or Older

The establishment of priority groups for vaccination is an important strategy, based on epidemiological indicators and the characterization of the vulnerability of the groups [50]. Thus, elderly citizens were the first to be vaccinated in Brazil, prioritized by age group.

However, for the second analysis, elderly people aged over 64 years were chosen to represent about 13% of the population [51].

The Euclidean distance was chosen as the dissimilarity measure, and for the composition of the clusters, the average method was used, which also presented better results, with a value of *CCC* = 0.889. The cut on the dendrogram dissimilarity axis was at the height of 1.8, which results in the agglomeration of eight probable groups, as can be seen in Figure 5.

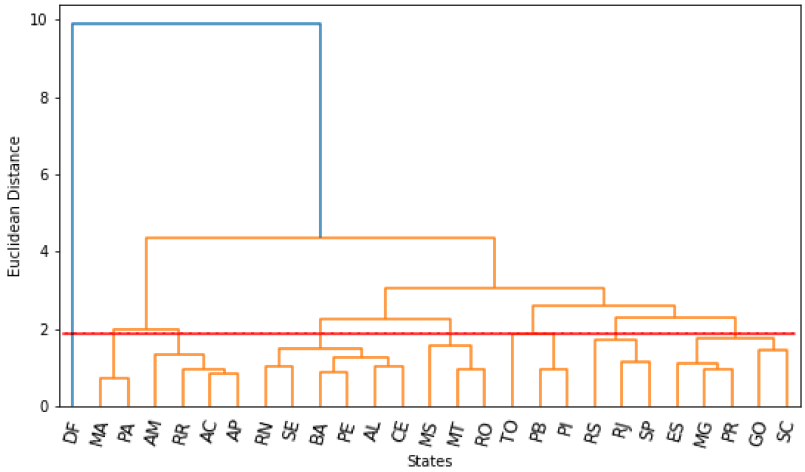

Figure 5. Dendrogram generated with data of the elderly 64 years or older. The multivariate analysis technique was applied to the 2021 data, grouping Brazil in a different format from the Brazilian regions, which can be observed in the branches of the tree. It presents satisfactory *CCC* (0.889), and eight clusters were suggested.

The first cluster, DF, a cluster with a single state, stands out isolated as it is the state with the highest average number of new cases per day (outlier) in just one state. The second group is composed of two states (MA and PA). The next cluster with the states AM, RR, AC, and AP has 14.81% representation, followed by the cluster with the highest number of states, six in all (22.22%) and from the same geographic region (RN, Sergipe (SE), Bahia (BA), Pernambuco (PE), Alagoas (AL) and CE). The fifth, sixth, and seventh clusters, with three states: Mato Grosso do Sul (MS), Mato Grosso (MT), and Rondônia (RO); TO, Paraíba (PB), and Piauí (PI); and Rio Grande do Sul (RS), Rio de Janeiro (RJ), and SP. The last cluster represents 18.53% of the states, with Espírito Santo (ES), Minas Gerais (MG), Paraná (PR), Goiás (GO), and Santa Catarina (SC).

Figure 6 presents the result of the elbow method for the database of elderly people aged 64 and over, suggesting the number of eight clusters.

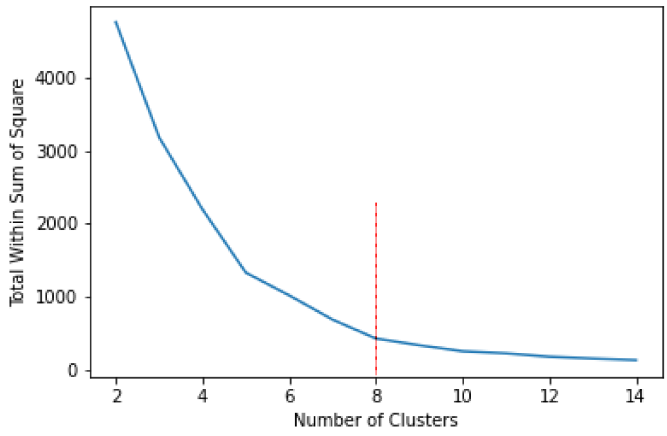

Figure 6. Curve of elbow method generated with data from the elderly 64 years or older. The number of the cluster was determined by looking at the point position on the "elbow" arm.

It can be seen that the second cluster analysis resulted in different groups from the previous analysis, as illustrated in Figure 5, which is better visualized on the Brazilian Map in Figure 7. More clusters were found with similar characteristics among themselves, and differences among the clusters.

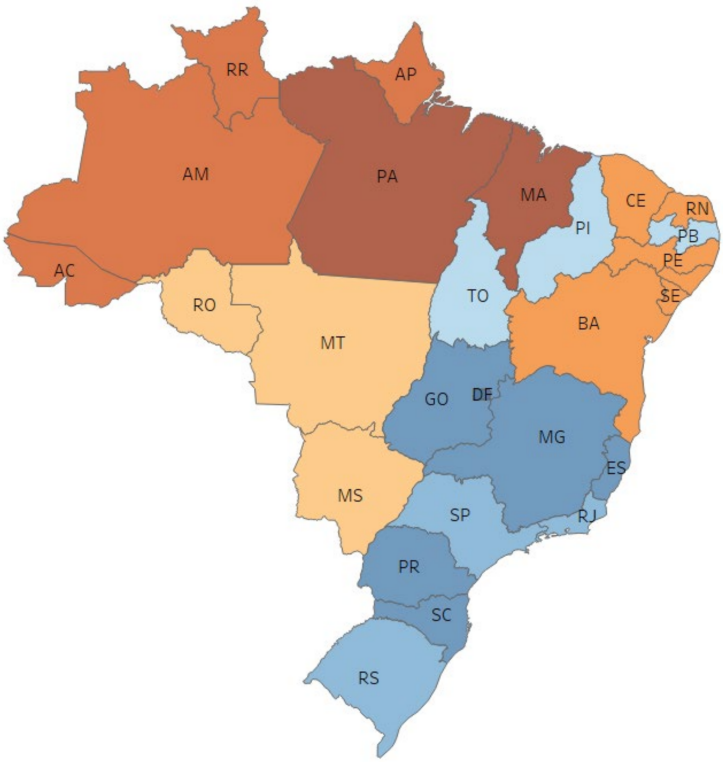

Figure 7. Map of Brazil generated from the result of the dendrogram with eight clusters (age group aged 64 and over). States well distributed in the clusters, showing heterogeneity in the behavior of the population.

This difference in divisions is very marked, in eight clusters tested with data from the elderly (13% of the population) to six clusters (all age groups) and with a different number of states per cluster, possibly due to factors such as:

- vaccination of the elderly was the first in the vaccination calendar (and continued throughout the year 2021) and had no results on the efficacy of vaccines, leading to mistrust [52];
- vaccination campaigns were starting without many disclosures and some calendars did not contain information on dates for each group and where to receive the doses. Each state had the autonomy to prepare the calendar and dissemination;
- vaccines were missing in many states and started applications through the capitals (sometimes not arriving in the interiors). According to [53], there were three errors that led to the lack of vaccines: the government did not anticipate and buy vaccines in 2020, there was a lack of definition on who should be vaccinated first, and a lack of training caused a waste of doses;
- vaccination infodemics were well-exposed and circulated on social media and the Internet, generating fear and vaccine hesitation. People began to discredit science, believe in fake news, and act against science [52].
- in addition, the elderly were identified as the most vulnerable in the dissemination of fake news [54], and they are seven times more likely to spread false news compared to people under 29 years [55]. This generated a pertinent concern because the presence of the elderly as internet users has been growing and this has been shown to be the largest proportional increase among the age groups [55].

The more infodemics shared, the greater the amount of fear and mistrust in the population. In this way, the patterns of behavior among the 27 federative units could be more heterogeneous.

5. Conclusions

The high demand for information corresponding to the growing popularity of COVID-19 vaccination research in news sources highlights the importance of public health officials working with the media to ensure that information is correct. This is because a high number of searches for vaccination infodemics can make it difficult for vaccination campaigns to be productive.

In this sense, the work used infodemic information on COVID-19 vaccination in Brazil, collected from the GT for the year 2021, with information on the number of vaccinated population and other important variables to perform a multivariate analysis of data and employ dendrograms to cluster the Brazilian states.

The use of the clustering technique, for the two analyses performed, resulted in six and eight clusters, respectively. Different results were found in the number of clusters and the aggregated states, composed of states with a high probability of having similar characteristics within each group and differences from the others. The results obtained with the *CCC* indicated a good degree of fit between the dendrograms and the dissimilarity matrices, allowing inferences to be made from the graphic representation. Finally, the UPGMA clustering algorithm was the most efficient, providing the lowest degrees of linkage and the highest *CCC* values.

In the analysis with vaccination data of the elderly aged 64 years or older, more heterogeneity in the patterns is visualized. During this period, the population had distrust and fear of vaccine efficacy, generating more sharing and infodemic research on vaccines. Vaccination in the states followed at different rates.

After analyzing the generated database, six clusters with more clustered states and more homogeneity was observed. The number of vaccinations increased in the age groups as clarifications were made about the importance and efficacy of the vaccine, leading to a significant decrease in both the number of cases and the number of deaths per day, and, of course, more vaccines were applied with improvements in the rate of evolution of pandemic numbers.

Future works should combine more information, expand data to two years of vaccination (2021 and 2022), and add other variables that express social and sociodemographic inequality, such as age and sex, which would provide more potential explanations for the behavior of the Brazilian population. It is intended to use the technique of Bayesian networks that would help policymakers and health managers understand which variables are most relevant.

Author Contributions: Conceptualization, M.d.P.H. and C.R.F.; Methodology, M.d.P.H. and C.R.F.; Software, M.d.P.H.; Validation, M.d.P.H. and M.S.; Formal analysis, M.d.P.H., M.S. and C.R.F.; Investigation, M.d.P.H. and C.R.F.; Data curation, M.d.P.H.; Writing−original draft, M.d.P.H.; Writing−review and editing, M.d.P.H., L.S., T.A., N.V. and C.R.F.; Visualization, L.S., T.A. and N.V.; Supervision, L.S., N.V. and C.R.F. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank the Coordination for the Improvement of Higher Education (CAPES) and Dean of Research and Graduate Programs of the Federal University of Para (PROPESP/UFPA) for their program supporting qualified publishing (PAPQ)-funding number 02/2022, which facilitated the funding that enabled the payment of publication fees.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on github, access link: https://github.com/mpenhaharb/InfodemicsVaccine (accessed on 25 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AC	Acre
AL	Alagoas
AM	Amazonas
AP	Amapá
BA	Bahia
CCC	Cophenetic Correlation Coefficient
CE	Ceará
DF	Distrito Federal
ES	Espírito Santo
GO	Goiás
GT	Google Trends
MA	Maranhão
MG	Minas Gerais
MS	Mato Grosso do Sul
MT	Mato Grosso
PA	Pará
PB	Paraiba
PE	Pernambuco
PI	Piauí
PR	Paraná
RJ	Rio de Janeiro
RN	Rio Grande do Norte
RO	Rondônia
RR	Roraima
RS	Rio Grande do Sul
SC	Santa Catarina
SD	Standard Deviation
SE	Sergipe
SP	São Paulo
TO	Tocantins
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
WHO	World Health Organization

References

1. Guo, Y.; Cao, Q.; Hong, Z.; Tan, Y.; Chen, S.; Jin, H.; Tan, K.; Wang, D.; Yan, Y. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—An update on the status. *Mil. Med. Res.* **2020**, *7*, 11. [CrossRef] [PubMed]
2. Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, P.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **2020**, *395*, 565–574. [CrossRef]
3. World Health Organization. Coronavirus Disease (COVID-19-2019) Situation Reports—51. Available online: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10 (accessed on 3 April 2022).
4. World Health Organization. Coronavirus Disease (COVID-19) Dashboard. Available online: https://covid19.who.int (accessed on 6 June 2022).
5. Kouzy, R.; Jaoude, J.A.; Kraitem, A.; Alam, M.B.E.; Karam, B.; Adib, E.; Zarka, J.; Traboulsi, C.; Akl, E.W.; Baddour, K. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus* **2020**, *12*, 3. [CrossRef] [PubMed]
6. Ghebreyesus, T.A. In Proceedings of the Munich Security Conference. Munich, Germany, 15 February 2022; Available online: https://www.who.int/director-general/speeches/detail/munich-security-conference (accessed on 15 March 2022).
7. Department of Global Communications. UN Tackles "Infodemic" of Misinformation and Cybercrime in COVID-19 Crisis. Available online: https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-COVID-19 (accessed on 15 March 2022).
8. Leem, M.; You, M. Direct and Indirect Associations of Media Use with COVID-19 Vaccine Hesitancy in South Korea: Cross-sectional Web-Based Survey. *J. Med. Internet Res.* **2022**, *24*, e32329. [CrossRef]
9. Batista, S.R.; de Souza, A.S.S.; Nogueira, J.; de Andrade, F.B.; Thumé, E.; Teixeira, D.S.d.C.; Lima-Costa, M.F.; Facchini, L.A.; Nunes, B.P. Comportamentos de proteção contra COVID-19 entre adultos e idosos brasileiros que vivem com multimorbidade: Iniciativa ELSI-COVID-19. *Cad. Saúde Pública* **2020**, *36*, e00196120. [CrossRef]

10. Organização Pan-Americana de Saúde. Dez Ameaças à Saúde Global em 2019. Available online: https://www.paho.org/pt/noticias/17-1-2019-dez-ameacas-saude-que-oms-combatera-em-2019 (accessed on 20 March 2022).
11. Sociedade Brasileira de Imunização. Especialistas se Reúnem para Debater o Fenômeno da Hesitação Vacinal no Brasil. Available online: https://sbim.org.br/noticias/1619-especialistas-se-reunem-para-debater-o-fenomeno-da-hesitacao-vacinal-no-brasil (accessed on 25 March 2022).
12. Google Noticias. Coronavírus (COVID_19). Available online: https://news.google.com/covid19/map?hl=pt-BR&gl=BR&ceid=BR%3Apt-419 (accessed on 5 June 2022).
13. Johns Hopking. Vaccination Progress across the World. Available online: https://coronavirus.jhu.edu/vaccines/international (accessed on 25 June 2022).
14. Nexo Jornal. Como Bolsonaro Atacou e Atrasou a Vacinação na Pandemia. Available online: https://www.nexojornal.com.br/expresso/2021/03/21/Como-Bolsonaro-atacou-e-atrasou-a-vacina%C3%A7%C3%A3o-na-pandemia (accessed on 10 June 2022).
15. Unicamp. Negacionismo na Pandemia: A Virulência da Ignorância. Available online: https://www.unicamp.br/unicamp/noticias/2021/04/14/negacionismo-na-pandemia-virulencia-da-ignorancia (accessed on 10 June 2022).
16. Wilson, K.; Brownstein, J.S. Early detection of disease outbreaks using the Internet. *Can. Méd. Assoc. J.* **2009**, *180*, 829–831. [CrossRef]
17. Arora, V.S.; Mckee, M.; Stuckler, D. Google Trends: Opportunities and limitations in health and health policy research. *Health Policy* **2019**, *123*, 338–341. [CrossRef]
18. Mangono, T.; Smittenaar, P.; Caplan, Y.; Huang, V.H.; Sutermaster, S.; Kemp, H.; Sgaier, S.H. Information-Seeking Patterns During the COVID-19 Pandemic Across the United States: Longitudinal Analysis of Google Trends Data. *J. Med. Internet Res.* **2021**, *23*, e22933. [CrossRef]
19. Rovetta, A.; Bhagavathula, A.S. COVID-19-Related Web search behaviors and infodemic attitudes in Italy: Infodemiological Study. *JMIR Public Health Surveill* **2020**, *6*, e19374. [CrossRef]
20. Ceron, W.; Sanseverino, G.G.; Santos, M.L.; Quiles, M.G. COVID-19 fake news diffusion across Latin America. *Soc. Netw. Anal. Min.* **2021**, *11*, 47. [CrossRef]
21. Custodio, M.; Peñaloza, R.; Alvarado, J.; Chanamé, F.; Maldonado, E. Surface Water Quality in the Mantaro River Watershed Assessed after the Cessation of Anthropogenic Activities Due to the COVID-19 Pandemic. *Pol. J. Environ. Stud.* **2021**, *30*, 3005–3018. [CrossRef]
22. Silva, R.M.; Mendes, C.F.; Manchein, C. Scrutinizing the heterogeneous spreading of COVID-19 outbreak in Brazilian territory. *Phys. Biol.* **2021**, *18*, 025002. [CrossRef]
23. Shafiq, A.; Çolak, A.B.; Sindhu, T.N.; Lone, S.A.; Alsubie, A.; Jarad, F. Comparative study of artificial neural network versus parametric method in COVID-19 data analysis. *Results Phys.* **2022**, *38*, 105613. [CrossRef]
24. Shafiq, A.; Sindhu, T.N.; Alotaibi, N. A novel extended model with versatile shaped failure rate: Statistical inference with F-19 applications. *Results Phys.* **2022**, *36*, 105398. [CrossRef]
25. James, N.; Menzies, M.; Bondell, H. Comparing the dynamics of COVID-19 infection and mortality in the United States, India, and Brazil. *Phys. D Nonlinear Phenom.* **2022**, *432*, 133158. [CrossRef]
26. James, N.; Menzies, M. COVID-19 in the United States: Trajectories and second surge behavior. *Chaos Interdiscip. J. Nonlinear Sci.* **2020**, *30*, 091102. [CrossRef]
27. Harb, M.A.; Silva, L.; Vijaykumar, N.L.; Silva, M.S.; Francês, C.R. An Analysis of the Deleterious Impact of the Infodemic during the COVID-19 Pandemic in Brazil: A Case Study Considering Possible Correlations with Socioeconomic Aspects of Brazilian Demography. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3208. [CrossRef]
28. Braz, A.M.; Oliveira, I.J.; Cavalcanti, L.C.; Almeida, A.C.; Chávez, E.S. Cluster Analysis for Landscape Typology. *Mercator* **2020**, *19*, e19011. [CrossRef]
29. OpenDataSUS. Registros de Vacinação COVID19. Available online: https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao/resource/5093679f-12c3-4d6b-b7bd-07694de54173 (accessed on 15 March 2022).
30. Painel Coronavírus. Dados COVID-19. Available online: https://covid.saude.gov.br/ (accessed on 16 March 2022).
31. Painéis de Dados ANATEL. Banda Larga Fixa. Available online: https://informacoes.anatel.gov.br/paineis/acessos/bandalarga-fixa (accessed on 15 March 2022).
32. Trends. Veja o Que o Mundo está pesquisando. Available online: https://trends.google.com.br/trends/?geo=BR (accessed on 22 March 2022).
33. Rovetta, A.; Bhagavathula, A.S. Global Infodemiology of COVID-19: Analysis of Google Web searches and instagram hashtags. *J. Med. Internet Res.* **2020**, *22*, e20673. [CrossRef]
34. Agência da Hora. Top 5 Fake News Mais Absurdas Sobre a Vacina. Available online: https://www.ufsm.br/midias/experimental/agencia-da-hora/2021/11/11/top-5-fake-news-mais-absurdas-sobre-a-vacina/ (accessed on 1 March 2022).
35. Diaz, L.C. The Lies That Are Told against Vaccines for COVID-19. Available online: https://www.revistaquestaodeciencia.com.br/artigo/2022/01/13/mentiras-que-se-contam-contra-vacinas-para-covid-19 (accessed on 4 March 2022).
36. Brasil de Fato. Você não vai se Transformar em Jacaré: 10 Mentiras Sobre Vacinas que Circulam por aí. Available online: https://www.brasildefato.com.br/2020/12/19/voce-nao-vai-se-transformar-em-jacare-10-mentiras-sobre-vacinas-que-circulam-por-ai (accessed on 2 March 2022).

37. Silva, R. De "Jacaré" a "Vacina do Doria": Relembre Frases de Bolsonaro Sobre Vacinação. Available online: https://www.agazeta.com.br/es/politica/de-jacare-a-vacina-do-doria-relembre-frases-de-bolsonaro-sobre-vacinacao-0121 (accessed on 5 April 2021).
38. Patel, V.R.; Mehta, R.G. Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm. *Int. J. Comput. Sci. Issues* **2011**, *8*, 331. Available online: https://www.ijcsi.org/articles/Impact-of-outlier-removal-and-normalization-approach-in-modified-kmeans-clustering-algorithm.php (accessed on 7 April 2022).
39. De Barros Vilela, G., Jr. Estatística: Teste Z (ou Escore Padronizado). Available online: http://www.cpaqv.org/estatistica/teste_z.pdf (accessed on 1 May 2022).
40. Fávero, L.L.; Belfiore, P.P.; Silva, F.L.; Chan, B.L. *Análise de Dados: Modelagem MULTIVARIADA para Tomada de Decisões*; Elsevier: Amsterdam, The Netherlands, 2009; ISBN 8535230467.
41. Metz, J. Interpretação de Clusters Gerados por Algoritmos de Clustering Hierárquico. Master's Thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP), São Carlos, Brazil, 2006. [CrossRef]
42. Machado, R.L. Desenvolvimento de um Algoritmo Imunológico para Agrupamento de Dados. Bachelor's Thesis, Universidade de Caxias do Sul, Caxias do Sul, Brazil, 2011. Available online: https://repositorio.ucs.br/handle/11338/1486 (accessed on 20 April 2022).
43. Berkhin, P. *Survey of Clustering Data Mining Techniques*; Accruel Software: San Jose, CA, USA, 2006; Available online: https://faculty.cc.gatech.edu/~{}isbell/classes/reading/papers/berkhin02survey.pdf (accessed on 1 May 2022).
44. Vicini, L. Análise Multivariada da Teoria à Prática. Available online: http://w3.ufsm.br/adriano/livro/Caderno%20dedatico%20multivariada%20-%20LIVRO%20FINAL%201.pdf (accessed on 20 April 2022).
45. Rohlf, F.J. Adaptative hierarquical clustering schemes. *Syst. Zool.* **1970**, *19*, 58–82. [CrossRef]
46. Kodinariya, T.M.; Makwana, P.R. Review on Determining Number of Cluster in K-Means Clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2013**, *1*, 2321–7782.
47. Syakur, M.A.; Khotimah, B.K.; Rochman, E.M.; Satoto, B.D. Integration K-Means clustering method and elbow method for identification of the best customer profile cluster. *Conf. Ser. Mater. Sci. Eng.* **2018**, *336*, 012017. Available online: https://iopscience.iop.org/article/10.1088/1757-899X/336/1/012017 (accessed on 20 April 2022). [CrossRef]
48. IBGE Educa. Uso de Internet, Televisão e Celular no Brasil. Available online: https://educa.ibge.gov.br/jovens/materias-especiais/20787-uso-de-internet-televisao-e-celular-no-brasil.html (accessed on 10 April 2022).
49. Fundação Oswaldo Cruz. COVID-19: Balanço de dois anos da Pandemia Aponta Vacinação como Prioridade. Available online: https://portal.fiocruz.br/noticia/covid-19-balanco-de-dois-anos-da-pandemia-aponta-vacinacao-como-prioridade (accessed on 11 April 2022).
50. Souto, E.P.; Kabad, J. Hesitação vacinal e os desafios para enfrentamento da pandemia de COVID-19 em idosos no Brasil. *Rev. Bras. Geriatr. Gerontol.* **2020**, *23*, e210032. [CrossRef]
51. IBGE População. Projeção da População do Brasil e das Unidades da Federação. Available online: https://www.ibge.gov.br/apps/populacao/projecao/index.html (accessed on 6 June 2022).
52. Agência da Hora. Por que a Vacinação Contra COVID-19 no Brasil não Segue o Ritmo de Campanhas Anteriores? Available online: https://www.ufsm.br/midias/experimental/agencia-da-hora/2021/05/10/por-que-a-vacinacao-contra-covid-19-no-brasil-nao-segue-o-ritmo-de-campanhas-anteriores/ (accessed on 6 June 2022).
53. BBC News. 3 Erros que Levaram à Falta de Vacinas Contra COVID-19 no Brasil. Available online: https://www.bbc.com/portuguese/brasil-56160026 (accessed on 6 June 2022).
54. Estabel, L.B.; Luce, B.F.; Santini, L.A. Idosos, fake news e letramento informacional. *Rev. Bras. Bibliotecon. Doc.* **2020**, *16*, 1–15. Available online: https://rbbd.febab.org.br/rbbd/article/view/1348/1206 (accessed on 2 June 2022).
55. Yabrude, A.Z.; Souza, A.M.; Campos, C.W.; Bohn, L.; Tiboni, M. Desafios das Fake News com Idosos durante Infodemia sobre COVID-19: Experiência de Estudantes de Medicina. *Rev. Bras. Educ. Med.* **2020**, *44*, e0140. [CrossRef]