Saneamento e Educação: Explorando Padrões em Municípios Brasileiros através de Clusterização

Pablo Henrique da Silva Lima[1]*; Miguel Ângelo Lellis Moreira[2]

[1] Universidade Federal Rural do Rio de Janeiro (UFRRJ). Bacharel em Administração.
[2] Universidade Federal Fluminense. Doutorando em Engenharia de Produção – Pesquisa Operacional.
R. Passo da Pátria, São Domingos, 24210-240, Niterói - RJ, Brasil.
*autor correspondente: lima.pablohs@gmail.com

Saneamento e Educação: Explorando Padrões em Municípios Brasileiros através de Clusterização

Resumo

O saneamento básico e a educação são pilares fundamentais para o desenvolvimento social e econômico, especialmente em países com grandes desigualdades regionais, como o Brasil. Apesar dos avanços nos últimos anos, milhões de brasileiros ainda carecem de acesso à água potável e coleta de esgoto, condições que afetam diretamente a saúde e o desempenho educacional. Este estudo investiga a relação entre saneamento básico e indicadores educacionais em municípios brasileiros, aplicando técnicas de análise de clusters aos dados do Censo de 2010 do IBGE. Utilizando o algoritmo K-means, os municípios foram agrupados em três clusters, com base em variáveis como acesso à água encanada, coleta de lixo e índices de escolaridade. Os resultados parciais mostram uma correlação positiva entre melhor infraestrutura sanitária e desempenho educacional superior, destacando que os municípios com melhores condições de saneamento apresentam os melhores índices educacionais. As conclusões sugerem a necessidade de políticas públicas integradas que abordem simultaneamente a melhoria do saneamento e da educação, com foco especial nos municípios mais vulneráveis.
Palavras-chave: Saneamento básico; Educação; Análise de clusters; Desigualdades regionais

Abstract

Basic sanitation and education are fundamental pillars for social and economic development, especially in countries with significant regional inequalities, such as Brazil. Despite recent advancements, millions of Brazilians still lack access to clean water and sewage systems, conditions that directly affect health and educational performance. This study investigates the relationship between basic sanitation and educational indicators in Brazilian municipalities by applying cluster analysis techniques to data from the 2010 IBGE Census. Using the K-means algorithm, municipalities were grouped into three clusters based on variables such as access to piped water, waste collection, and education levels. Preliminary results show a positive correlation between better sanitation infrastructure and superior educational performance, highlighting that municipalities with better sanitary conditions also have higher educational indices. The conclusions suggest the need for integrated public policies that simultaneously address improvements in sanitation and education, with a special focus on the most vulnerable municipalities.
Keywords: Basic sanitation; Education; Cluster analysis; Regional inequalities

Introdução

Segundo o artigo 205 da Constituição Federal do Brasil (1988), a educação é um direito de todos e um dever tanto do Estado quanto da família. Entretanto, conforme o relatório da Organização para a Cooperação e Desenvolvimento Econômico (OECD, 2020), o sistema educacional brasileiro continua apresentando baixo desempenho ao longo dos anos. Essa constatação suscita uma reflexão crítica sobre a efetividade dos investimentos diretos em educação, questionando se tais aportes são suficientes para promover melhorias significativas no desempenho escolar.

Segundo a Proposta de Emenda à Constituição nº 2, de 2016 (BRASIL, 2016), o saneamento básico é um direito social garantido pela Constituição, assim como a saúde, a alimentação, a educação, o trabalho, o lazer, o transporte e a moradia. No entanto, uma pesquisa realizada pelo Instituto Trata Brasil (2023) revela que aproximadamente 35 milhões de brasileiros ainda não têm acesso à água tratada, e cerca de 100 milhões carecem de serviços adequados de coleta de esgoto.

De acordo com a Lei nº 14.026, de 15 de julho de 2020 (BRASIL, 2020), conhecida como Marco Legal do Saneamento Básico, a universalização do acesso ao saneamento é um objetivo central da legislação. A meta é assegurar que 99% da população tenha acesso à água potável e 90% ao tratamento de esgoto até 2033. Além disso, a lei estabelece diretrizes para melhorar a eficiência na distribuição de água tratada, promover a racionalização do consumo e buscar a eficiência energética nos processos de tratamento de água e esgoto.

Segundo Abanyie et al. (2021), o investimento em WASH (água, saneamento e higiene) em escolas é essencial para criar um ambiente de aprendizagem propício, reduzir o absentismo e melhorar os resultados acadêmicos gerais dos alunos, trazendo assim uma abordagem de que apesar do investimento direto em educação ser essencial, não é suficiente por si só. Segundo Sharma et al. (2024), a melhoria nas instalações de água, saneamento e higiene está associada a conquistas educacionais mais elevadas. Isso se deve ao fato de que um ambiente escolar saudável facilita a concentração dos alunos em suas atividades, o que pode levar a um desempenho acadêmico superior.

Conforme discutido por Valencio, Valencio e Baptista (2023), a ocorrência de desastres está profundamente interligada ao saneamento básico e à pobreza. No entanto, ao analisar a literatura existente, percebe-se que há uma maior ênfase em estudos focados em aspectos socioeconômicos em comparação com aqueles que abordam saneamento básico.

O objetivo desta pesquisa é identificar e analisar padrões de saneamento básico e indicadores educacionais em diferentes municípios brasileiros, utilizando técnicas de análise de clusters. Conforme descrito por Prakash (2022), a análise de cluster é uma técnica comum de análise exploratória de dados, usada para entender a estrutura de um conjunto de dados. O algoritmo K-means, por exemplo, é amplamente utilizado para segmentar os dados em subgrupos distintos e não sobrepostos, sendo aplicado em diversas áreas, como segmentação de mercado, agrupamento de documentos e segmentação de imagens. Com isso, busca-se compreender como as variáveis de saneamento influenciam os resultados educacionais, destacando as disparidades regionais e apontando possíveis correlações entre infraestrutura sanitária e desempenho acadêmico. A pesquisa pretende fornecer insights que possam orientar políticas públicas voltadas para a melhoria das condições de saneamento e educação, contribuindo para a redução das desigualdades socioeconômicas no país.

Material e Métodos

A desigualdade no acesso a serviços básicos, como saneamento e educação, representa um grande desafio para o desenvolvimento socioeconômico brasileiro. Diversos estudos mostram que a carência de infraestrutura afeta diretamente o desempenho educacional e, consequentemente, o potencial de desenvolvimento das regiões mais vulneráveis. Nesse contexto, torna-se essencial investigar a relação entre esses dois fatores para orientar políticas públicas mais eficazes.

Para entender melhor essa dinâmica, a presente pesquisa adota técnicas de análise de clusters, que são amplamente utilizadas em estudos exploratórios de grandes volumes de dados. O método de clusterização K-means foi escolhido por sua capacidade de agrupar os municípios brasileiros em grupos homogêneos com base em variáveis de saneamento e educação, facilitando a identificação de padrões entre municípios.

O estudo se fundamentou através de análise de dados das seguintes variáveis de saneamento básico:

- População com acesso à água encanada
- População com banheiro em casa e acesso à água encanada
- População com acesso à coleta de lixo

E de educação:

- IDHM de Frequência escolar (proporção de crianças e jovens até 18 anos que estão frequentando a escola, considerando a população total dessa faixa etária)
- IDHM de Escolaridade (percentual de pessoas com 18 anos ou mais que concluíram pelo menos o ensino fundamental)
- Atraso escolar no ensino fundamental (percentual de estudantes que possuem idade superior à esperada para o ano escolar que estão cursando no ensino fundamental)

A pesquisa utilizou dados públicos sobre saneamento básico e taxas de rendimento escolar do Censo de 2010, divulgados pelo Instituto Brasileiro de Geografia e Estatística (2010), conforme apresentado na Tabela 1. Optou-se por utilizar dados de um Censo anterior, de 2010, pois o mais recente, realizado em 2022, ocorreu após a pandemia de coronavírus, que teve grande impacto na frequência escolar.

Tabela 1. Dados brutos do Censo 2010 por município
Fonte: IBGE, Censo 2010.

Territorialidades	% da população em domicílios com água encanada 2010	% da população que vive em domicílios com banheiro e água encanada 2010	% de pessoas em domicílios urbanos com coleta de lixo 2010	Subíndice de frequência escolar - IDHM Educação 2010	Subíndice de escolaridade - IDHM Educação 2010	% de 6 a 14 anos no ensino fundamental com 2 anos ou mais de atraso idade-série 2010	Renda per capita média do 4º quinto mais pobre 2010
Abadia de Goiás (GO)	93,06	99,01	99,83	0,702	0,489	22,55	611,3
Abadia dos Dourados (MG)	88,5	98,18	98,03	0,673	0,394	20,39	577,33
Abadiânia (GO)	94,5	94,7	98,49	0,669	0,433	14,92	536,24
Abaeté (MG)	98,4	97,43	98,42	0,689	0,363	15,83	567,51
Abaetetuba (PA)	68,86	44,71	97,86	0,599	0,432	24,47	295,64
Xique-Xique (BA)	84,97	74,2	93,94	0,547	0,368	29,67	270,66
Zabelê (PB)	80,03	80,38	99,75	0,755	0,355	6,78	328,83
Zacarias (SP)	96,25	100	100	0,802	0,476	2,9	635,81
Zé Doca (MA)	89,28	46,12	89,38	0,589	0,372	18,7	277,4
Zortéa (SC)	96,08	99,91	99,58	0,734	0,536	11,58	876,79

A partir destes dados, foram utilizadas técnicas de limpeza e transformações de dados com o uso da biblioteca Python de código aberto Pandas (McKinney, 2010). A qual, segundo

McKinney (2018), é projetada com o intuito de trabalhar com dados relacionais de forma fácil e ágil.

Para a organização dos dados, foi utilizada a função iloc do pacote pandas (versão 2.2.2), que permitiu limitar as linhas do banco de dados para análise dos municípios. A conversão de dados não numéricos para string foi realizada com o método astype. A transformação de valores percentuais para decimais foi obtida ao dividir as colunas por 100, com a limitação das casas decimais em 6, por meio da função round. Além disso, as linhas contendo dados faltantes foram removidas utilizando a função dropna.

Com os dados organizados por município, foi realizada a normalização dos dados através do *Z-score*. Larson e Farber (2015) definem o Z-score como o número de desvios padrão em que um valor 'X' é encontrado a partir da média μ, conforme representado na equação 1.

$$z = \frac{(x-\mu)}{\sigma} \quad (1)$$

Onde:

- z é o valor normalizado;
- x é o valor da variável original;
- μ é a média da variável;
- σ é o desvio padrão da variável.

Dentro do ambiente Python, esse processo foi realizado pela função *StandardScaler,* a qual está presente no pacote *Scikit-learn* (Cournapeau, 2010). De acordo com Reitz e Schlusser (2017), o *Scikit-learn* é uma biblioteca de aprendizado de máquina que oferece uma variedade de recursos, incluindo redução de dimensões, imputação de dados ausentes, modelos de regressão e classificação, modelos de árvore, agrupamento, ajuste automático de parâmetros do modelo e muito mais.

Conforme discutido por Aggarwal e Reddy (2013), o problema de clusterização de dados é amplamente estudado na literatura de mineração de dados e aprendizado de máquina devido às suas inúmeras aplicações, como a segmentação de mercado, filtragem colaborativa e análise de redes sociais. A clusterização, em sua essência, visa particionar pontos de dados em grupos o mais semelhantes possível, embora a definição exata do problema varie de acordo com o modelo adotado, como abordagens baseadas em distância ou em modelos probabilísticos. Além disso, a clusterização também pode atuar como um passo intermediário em outros problemas de mineração de dados, como a classificação e a detecção de outliers.

Segundo Bruce e Bruce (2019) a técnica de agrupamento (cluster) é utilizada para organizar dados em grupos distintos, baseando-se na similaridade entre os registros dentro

de cada grupo. O propósito dessa técnica é o reconhecimento de conjuntos de dados que sejam tanto significativos quanto relevantes. Esses conjuntos identificados podem ser empregados de maneira direta, submetidos a análises mais minuciosas ou utilizados como característica ou resultado em modelos de regressão ou classificação. Bruce e Bruce (2019) explicam que o método K-médias, considerado como o precursor dos métodos de agrupamento, continua sendo amplamente aplicado devido à sua simplicidade de algoritmo e capacidade de lidar com grandes volumes de dados.

Fávero e Belfiore (2024) ressaltam a diversidade de procedimentos existentes para a realização de uma análise de agrupamentos, destacando que a escolha de medidas de distância ou semelhança depende do tipo de variáveis em estudo, sejam elas métricas ou binárias. Além disso, uma vez escolhida a medida, o pesquisador deve decidir entre vários métodos de aglomeração, que podem ser hierárquicos ou não hierárquicos. O processo, que pode parecer simples ao buscar agrupar observações em clusters internamente homogêneos, revela-se complexo devido à variedade de combinações possíveis entre medidas e métodos de aglomeração. Portanto, é crucial que o pesquisador estabeleça critérios claros, fundamentados na teoria e nos objetivos da pesquisa, além de sua própria experiência e intuição, para alocar as observações nos grupos apropriados.

O algoritmo escolhido para a criação dos Clusters foi o K-means. De acordo com Fávero e Belfiore (2024), os esquemas de aglomeração não hierárquicos, como o procedimento k-means ou k-médias, são métodos em que os centros de aglomeração são definidos e as observações são alocadas a eles com base em sua proximidade. Diferentemente dos métodos hierárquicos, nos quais o pesquisador pode explorar várias possibilidades de alocação das observações e definir o número de clusters em cada estágio de agrupamento, os métodos não hierárquicos requerem a estipulação prévia da quantidade de clusters. A função objetivo do K-means é representada pela equação (2):

$$J = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2 \quad (2)$$

Onde:

- J é o valor total da função objetivo;
- k é o número de clusters;
- Cj é o conjunto de pontos de dados no cluster j;
- xi é um ponto de dados no cluster j;
- μ_j é o centroide do cluster j;
- $||x_i - \mu_j||^2$ é a distância euclidiana ao quadrado entre o ponto xi e μ_j.

Com os dados normalizados e o algoritmo definido, foi necessário determinar o número ideal de clusters. Segundo Géron (2019), uma abordagem precisa para selecionar o número ideal de clusters é o uso do método da silhueta, que é a média do coeficiente de silhueta sobre todas as instâncias. Este coeficiente é calculado através da equação 3, onde a diferença entre a distância média para o cluster mais próximo *(b)* e a distância média para as outras instâncias no mesmo cluster *(a)*, dividido pelo máximo entre *(a)* e *(b)*. Essa métrica varia entre -1 e +1, onde valores próximos de +1 indicam instâncias bem dentro de seu próprio cluster, valores próximos de 0 indicam instâncias próximas aos limites do cluster e valores próximos de −1 indicam possíveis atribuições incorretas de cluster.

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (3)$$

Onde:

- *a(i)* é a distância média entre o ponto *i* e todos os pontos dentro do mesmo cluster;
- *b(i)* é a distância média entre o ponto *i* e todos os pontos do cluster mais próximo ao qual *i* não pertence.

A Figura 1 apresenta os valores médios do coeficiente de silhueta para diferentes quantidades de clusters, auxiliando na escolha do número mais adequado para a análise.

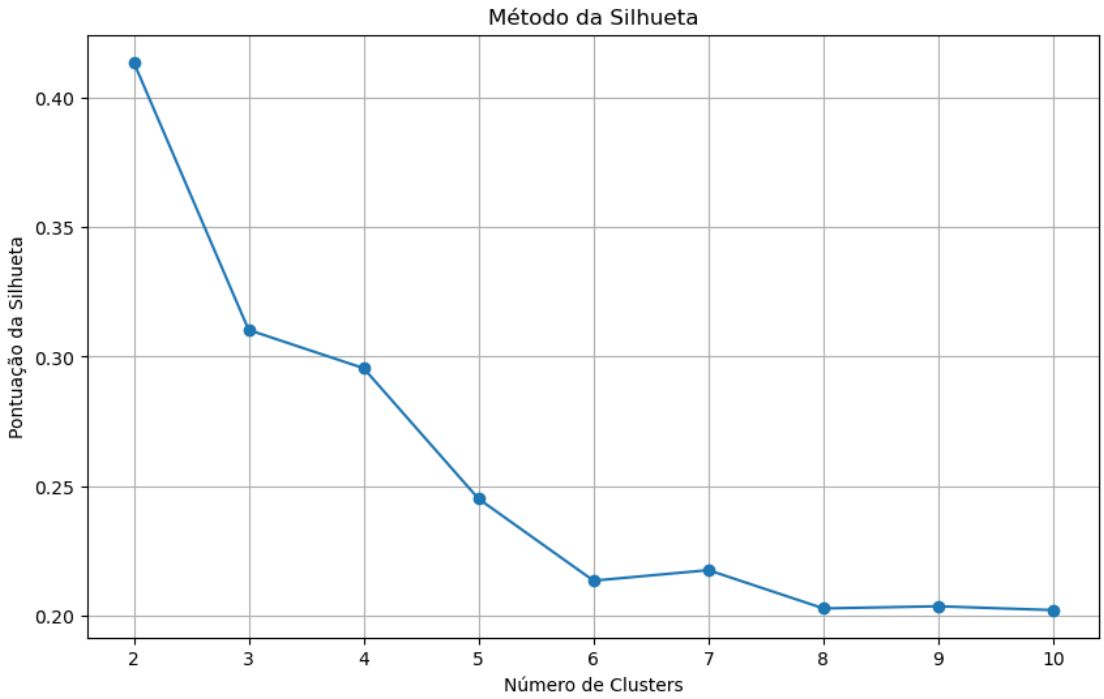

Figura 1. Método da silhueta
Fonte: Elaborado pelo autor

Outro método amplamente utilizado para determinar o número ideal de clusters é o Método do Cotovelo. Segundo Nainggolan et al. (2019), o método envolve a análise da variação do Erro Quadrático Total (SSE) com diferentes valores de k, conforme representado pela equação 4. Com isso, identifica-se o ponto onde a taxa de diminuição do SSE se torna menos acentuada, formando um "cotovelo" no gráfico. A escolha do número de clusters é baseada na localização deste ponto de inflexão, onde a redução do SSE começa a desacelerar significativamente.

$$\mathrm{SSE}(k) = \sum_{i=1}^{n} \sum_{j=1}^{k} 1_{\{x_i \in C_j\}} ||x_i - \mu_j||^2 \qquad (4)$$

Onde:

- $\mathrm{SSE}(k)$ é a soma dos erros quadráticos para o número de clusters k;
- x_i é o ponto de dados i;
- C_j é o conjunto de pontos de dados no cluster j;
- $||x_i - \mu_j||^2$ é a distância euclidiana ao quadrado entre o ponto x_i e μ_j.

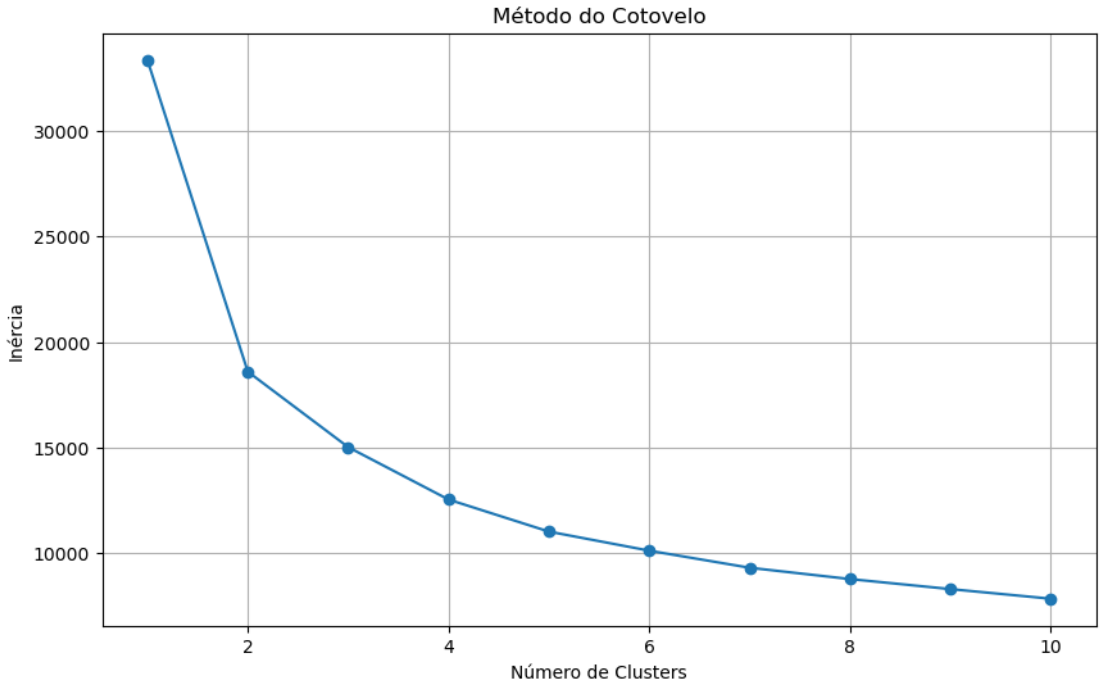

Figura 2. Método do Cotovelo
Fonte: Elaborado pelo autor

Sendo definido o número de clusters em 3, o algoritmo K-means pôde ser executado através da biblioteca Python Scikit-Learn (versão 1.5), assim agrupando os municípios em diferentes clusters, trazendo uma abordagem coletiva da análise.

Além do uso do algoritmo K-means para a formação de clusters, foi aplicada a correlação de Pearson para analisar a relação linear entre as variáveis selecionadas nos municípios. A correlação de Pearson é uma técnica estatística que quantifica o grau de associação linear entre duas variáveis métricas, com um valor de -1 indicando uma correlação negativa perfeita, 0 indicando nenhuma correlação linear, e 1 indicando uma correlação positiva perfeita (Montgomery & Runger, 2014).

Neste estudo, a correlação de Pearson foi utilizada para identificar quais variáveis estão fortemente relacionadas entre si, o que pode ajudar a justificar a escolha das variáveis para os clusters. Valores de correlação próximos a 1 ou -1 podem sugerir que variáveis relacionadas possam ser combinadas em um único fator, enquanto valores próximos a 0 indicam uma relação fraca, o que poderia justificar a necessidade de considerar variáveis distintas no processo de agrupamento. A correlação de Pearson pode ser calculada através da equação 5:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \quad (5)$$

Onde:

- r é o coeficiente de correlação;
- n é o número de observações;
- $\sum XY$ é a soma do produto de cada par de observações correspondentes das duas variáveis;
- $\sum X$ é a soma das observações da primeira variável;
- $\sum Y$ é a soma das observações da segunda variável;
- $\sum X^2$ é a soma dos quadrados das observações da primeira variável;
- $\sum Y^2$ é a soma dos quadrados das observações da segunda variável.

Para a criação de gráficos e trazer uma melhor visibilidade do resultado dos clusters, foi utilizada as bibliotecas Seaborn (Waskom, 2012) e Matplotlib (Hunter, 2002). Segundo McKinney (2018), a geração de visualizações informativas é uma etapa crucial na análise de dados, podendo ser parte do processo exploratório para identificar outliers ou necessidades de transformações nos dados, além de ser uma forma de gerar insights para modelos.

Resultados e Discussão

Após a remoção dos municípios sem dados, foram totalizados 5.556 municípios, distribuídos entre os clusters da seguinte forma:

- Cluster 0: 2.018 municípios
- Cluster 1: 860 municípios
- Cluster 2: 2.678 municípios

A distribuição dos municípios nos diferentes clusters pode ser visualizada na Figura 3.

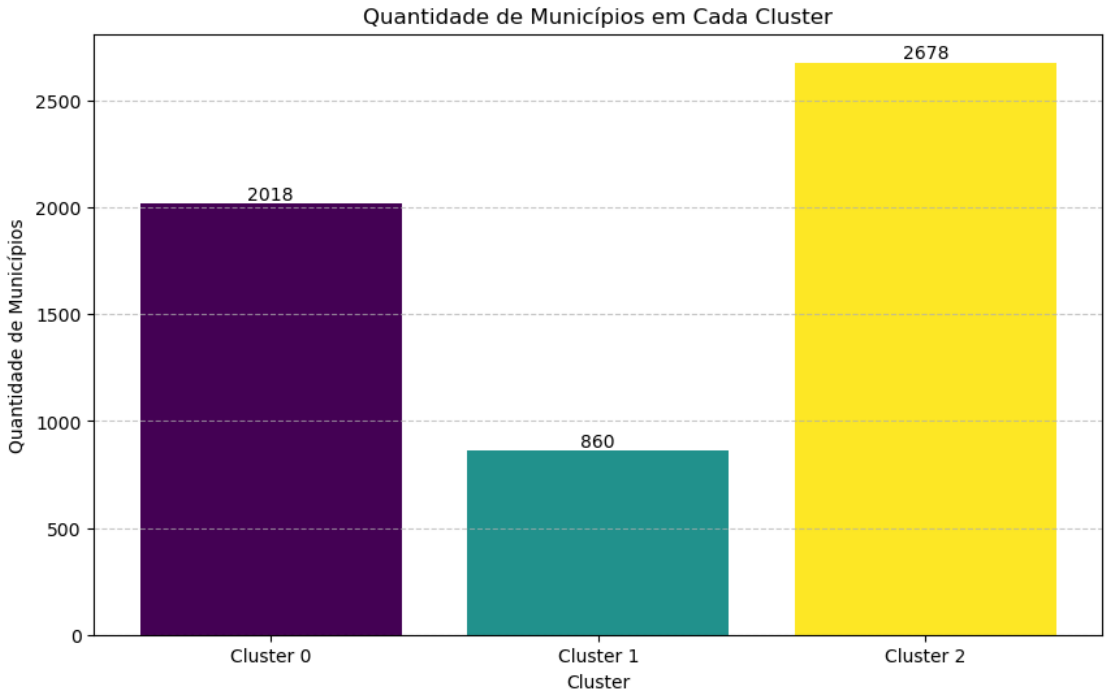

Figura 3. Divisão de municípios por cluster
Fonte: Elaborado pelo autor

A Figura 4 apresenta a distribuição geográfica dos municípios brasileiros agrupados pelos 3 Clusters. Observa-se que o Cluster 2 predomina nas regiões Sul e Sudeste, sugerindo uma maior homogeneidade socioeconômica. Por outro lado, as regiões Norte e Nordeste concentram uma proporção significativa dos Clusters 0 e 1, indicando desigualdades regionais mais marcantes e potenciais desafios estruturais nesses locais. Essa distribuição reforça a necessidade de políticas regionalizadas para atender às necessidades específicas de cada cluster.

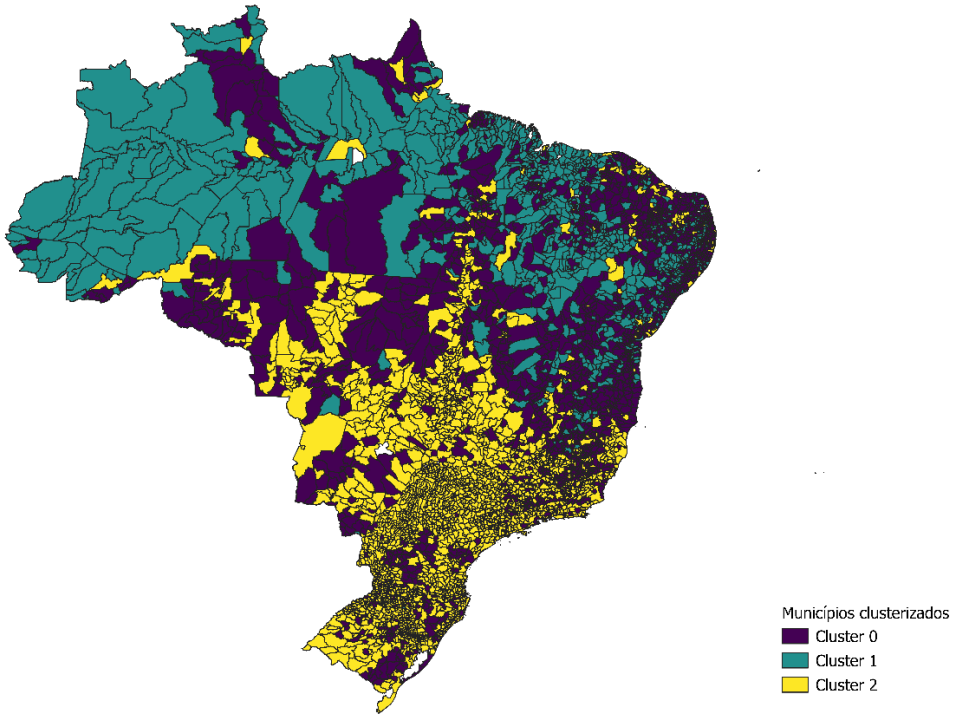

Figura 4. Mapa do Brasil com os Clusters
Fonte: Elaborado pelo autor

Segundo Bação, Mazeron e Simões (2023), a desigualdade no Brasil permanece entre as mais altas do mundo, com diferenças significativas observadas entre regiões e municípios, o que reflete disparidades socioeconômicas históricas que continuam a afetar a distribuição de riqueza e oportunidades no país.

Na Figura 5, é apresentada a quantidade de municípios em cada Cluster por região. O Cluster 0 é amplamente predominante no Nordeste. O Cluster 1 está mais concentrado na região Norte. Por sua vez, o Cluster 2 destaca-se no Sul e Sudeste. Essa distribuição evidencia as disparidades entre as regiões e reforça a importância de soluções contextualizadas para promover equilíbrio no desenvolvimento nacional.

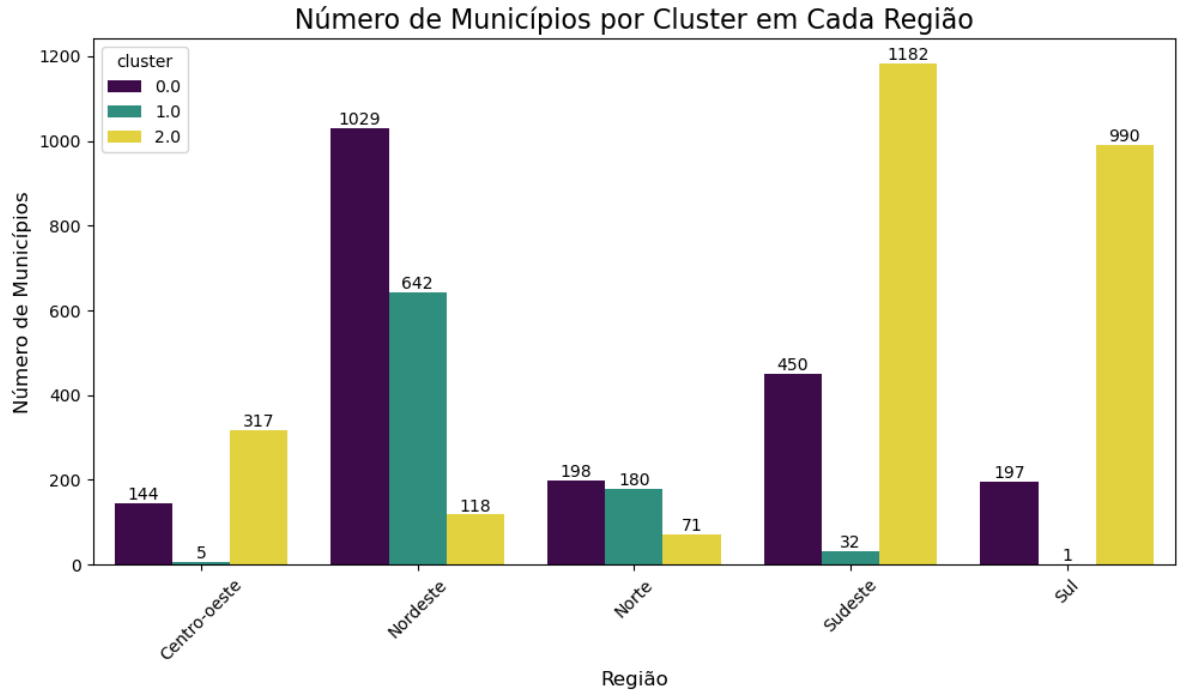

Figura 5. Gráfico de barras com o número de Municípios em cada Cluster por Região
Fonte: Elaborado pelo autor

Os municípios se agruparem em clusters com números tão distintos aponta a desigualdade regional no Brasil. Para identificar o padrão de cada cluster, foram gerados Boxplots de cada variável, para assim entender como estão postulados os municípios de cada cluster.

O Boxplot da Figura 6 ilustra a distribuição da população com acesso à água encanada em 2010 para os três clusters identificados. Observa-se que o Cluster 2 concentra a maioria dos municípios com acesso elevado à água encanada, apresentando menores variações. Em contraste, o Cluster 1 representa os municípios com menor acesso, exibindo uma distribuição mais ampla e índices mais baixos. O Cluster 0 ocupa uma posição intermediária, com uma variabilidade considerável entre os municípios. Esses dados reforçam as disparidades regionais no acesso a serviços básicos.

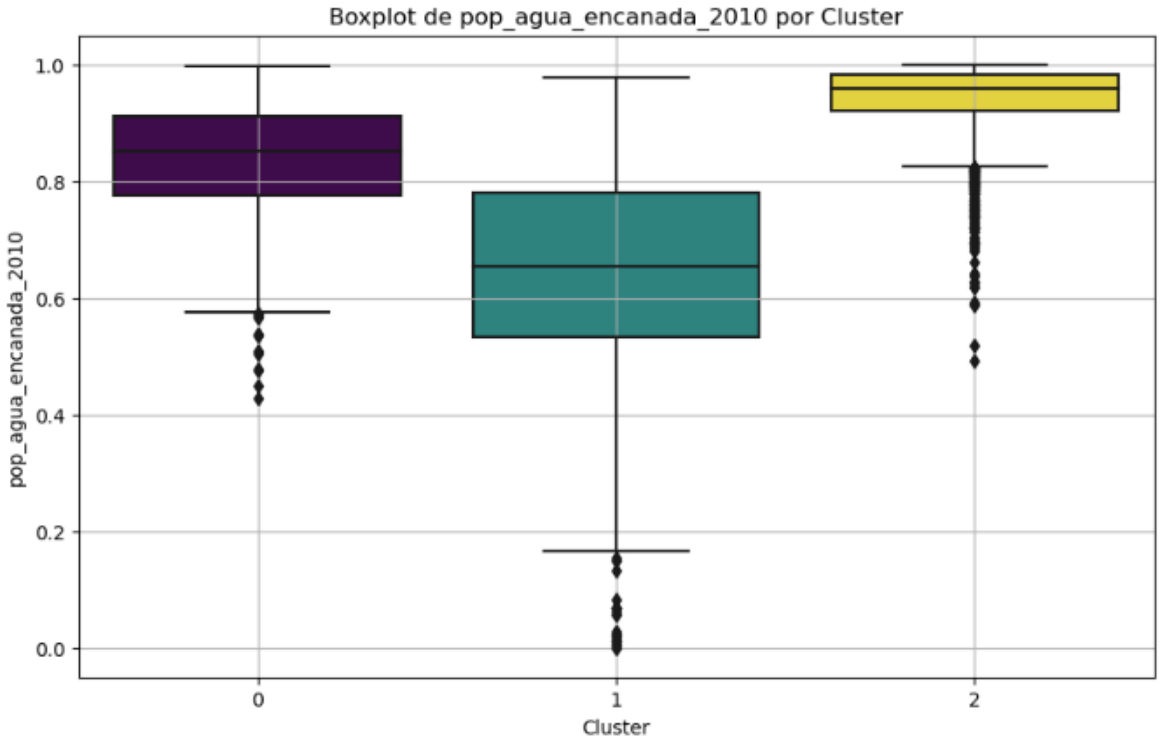

Figura 6. Boxplot de população com acesso à água encanada em 2010 por Cluster
Fonte: Elaborado pelo autor

A Figura 7 apresenta a distribuição da população com acesso tanto à água encanada quanto a banheiros em 2010. O Cluster 2 mantém um perfil semelhante ao anterior, com a maioria dos municípios apresentando altos níveis de acesso e pouca variação. O Cluster 1, novamente, destaca-se pelos menores índices e maior dispersão, indicando uma precariedade significativa em relação a esses serviços essenciais. Já o Cluster 0 ocupa uma posição de intermediária, com indicadores mais distribuídos, refletindo a diversidade de condições nos municípios desse grupo.

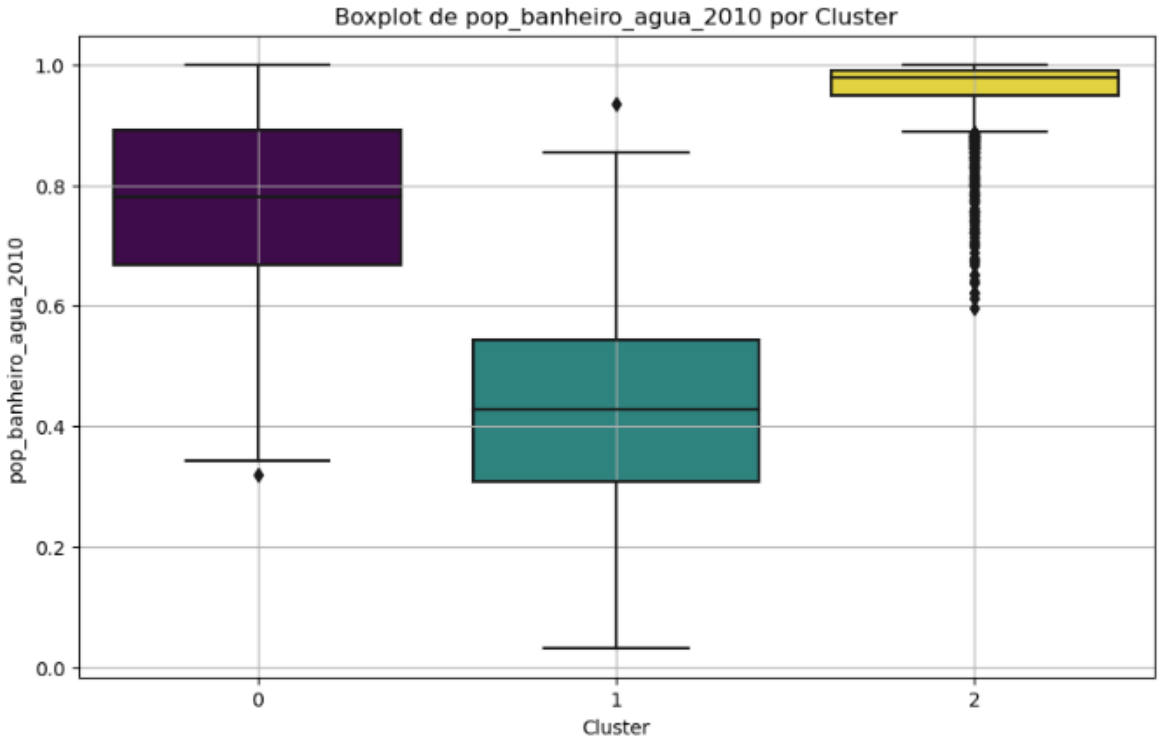

Figura 7. Boxplot de população com acesso à água encanada e banheiro em 2010 por Cluster
Fonte: Elaborado pelo autor

A Figura 8 revela a distribuição da população com acesso à coleta de lixo por cluster. O Cluster 2 inclui a maioria dos municípios com coleta de lixo regular, apresentando valores elevados e uma variação relativamente baixa. O Cluster 1, por outro lado, engloba os municípios com menor acesso à coleta de lixo, reforçando as dificuldades enfrentadas em áreas com infraestrutura sanitária insuficiente. O Cluster 0 mostra uma distribuição mais variada, com alguns municípios tendo acesso adequado à coleta de lixo e outros enfrentando dificuldades.

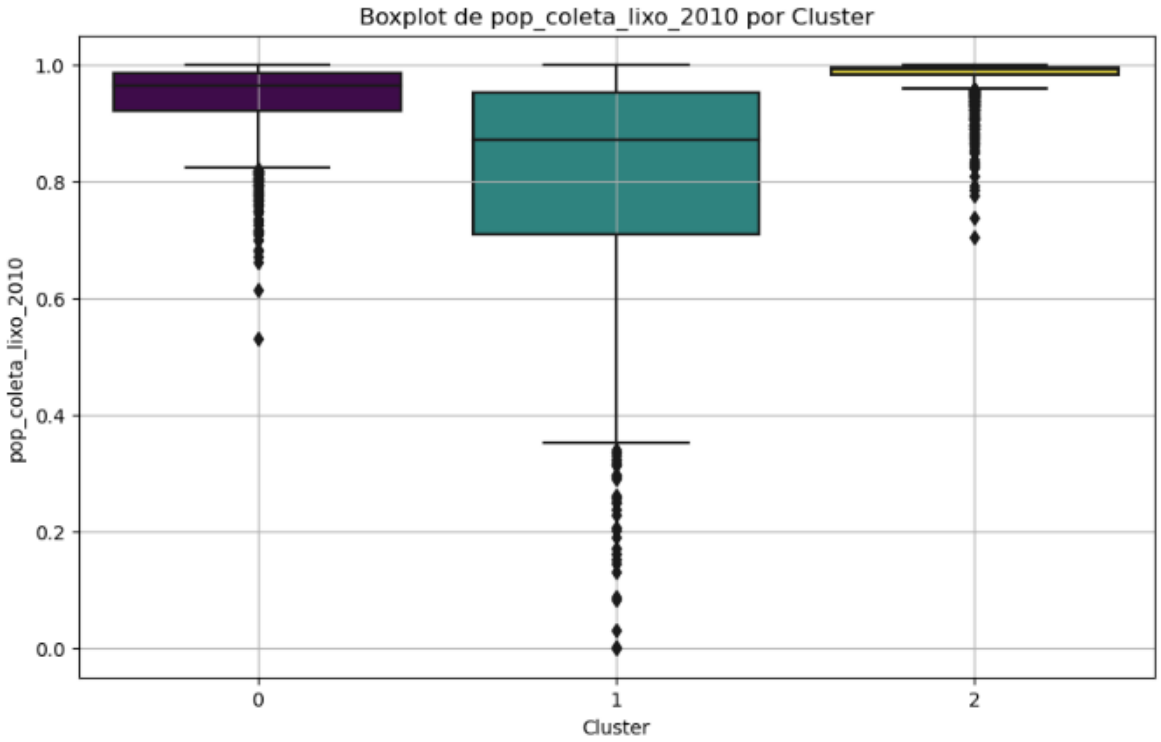

Figura 8. Boxplot de população com acesso à coleta de lixo em 2010 por Cluster
Fonte: Elaborado pelo autor

A Figura 9 ilustra a distribuição da frequência escolar de jovens até 18 anos. O Cluster 2 agrupa municípios com os melhores índices de frequência escolar, sugerindo uma correlação entre infraestrutura básica e presença dos alunos nas escolas. O Cluster 1, por outro lado, possui as menores taxas de frequência, com maior variação, refletindo as dificuldades educacionais enfrentadas em regiões com menor acesso a saneamento. O Cluster 0 apresenta um perfil intermediário, com uma ampla dispersão dos índices.

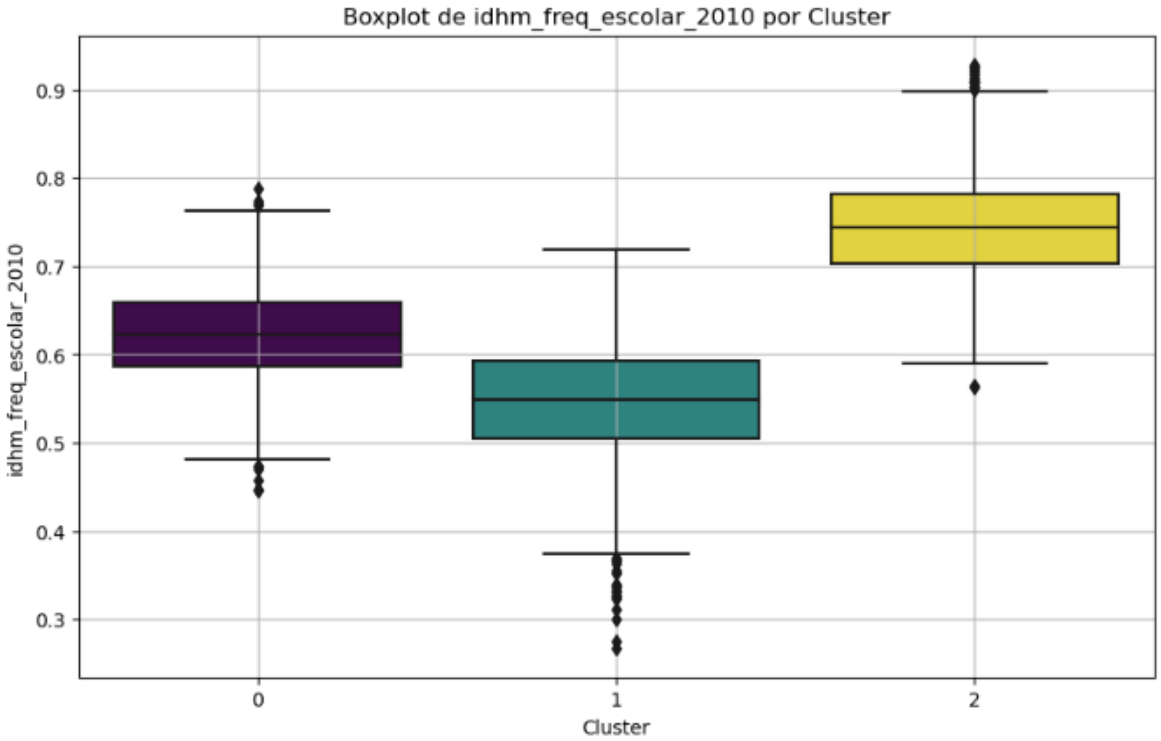

Figura 9. Boxplot de frequência escolar de jovens até 18 anos por Cluster
Fonte: Elaborado pelo autor

A Figura 10 mostra a porcentagem da população de mais de 18 anos com ensino fundamental completo em 2010. Os municípios do Cluster 2 se destacam por apresentar os maiores percentuais de escolaridade, com variação reduzida, enquanto o Cluster 1 mostra os índices mais baixos e dispersos, evidenciando as desigualdades educacionais. O Cluster 0, mais uma vez, ocupa uma posição intermediária, com uma maior variabilidade.

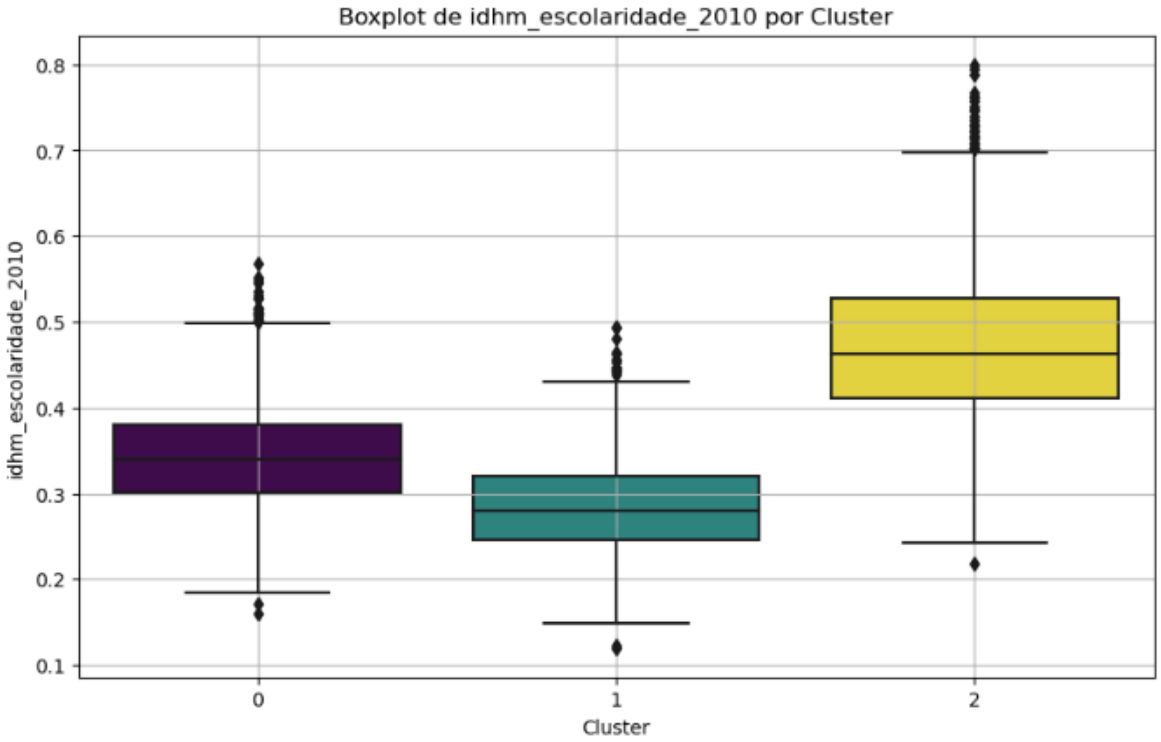

Figura 10. Boxplot de população de mais de 18 anos com ensino fundamental completo em 2010 por Cluster
Fonte: Elaborado pelo autor

A Figura 11 mostra a distribuição do atraso escolar no ensino fundamental em 2010. O Cluster 2 apresenta os menores índices de atraso, indicando uma menor incidência de defasagem idade-série. O Cluster 1, por outro lado, destaca-se pelos maiores níveis de atraso escolar, com uma ampla variação, refletindo os desafios enfrentados pelos municípios mais vulneráveis. O Cluster 0 ocupa uma posição intermediária, com níveis de atraso mais dispersos.

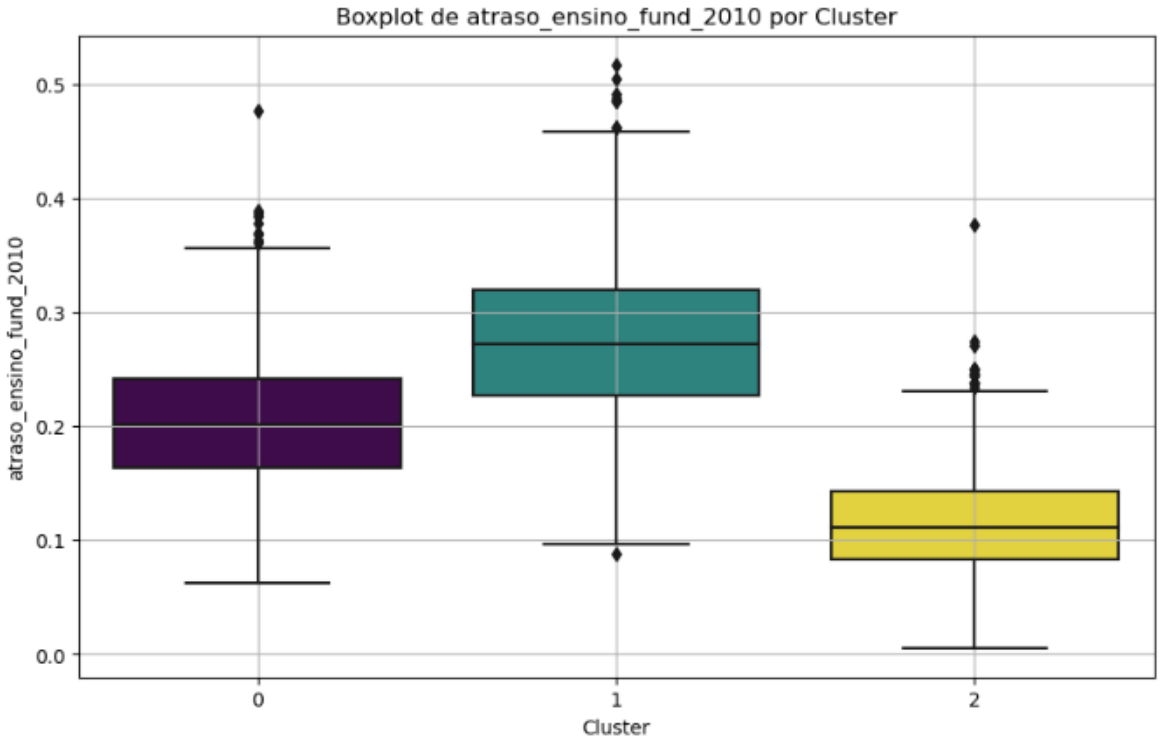

Figura 11. Boxplot de atraso escolar no ensino fundamental em 2010 por Cluster
Fonte: Elaborado pelo autor

A partir dos Boxplots, é possível observar que o Cluster 2 representa os municípios mais desenvolvidos, caracterizados por altos índices de acesso à água encanada e elevados níveis de escolaridade. O Cluster 1 agrupa os municípios em condições mais precárias, com acesso limitado à água encanada e baixos indicadores educacionais. Já o Cluster 0 inclui os municípios em uma situação intermediária, apresentando indicadores moderados tanto em saneamento quanto em educação.

A análise das disparidades entre os municípios brasileiros tornou-se evidente ao agrupar os municípios em clusters com base em variáveis como o acesso à água encanada e acesso à educação de base, como ilustra a Figura 12. Essa organização permitiu identificar com clareza as diferenças significativas nos níveis de infraestrutura básica e no desenvolvimento educacional entre as regiões. Municípios com melhor infraestrutura de saneamento tendem a apresentar também indicadores educacionais mais elevados, enquanto aqueles com acesso limitado a esses serviços estão associados a menores níveis de escolaridade, reforçando a interdependência entre saneamento e educação como fatores determinantes para o desenvolvimento socioeconômico.

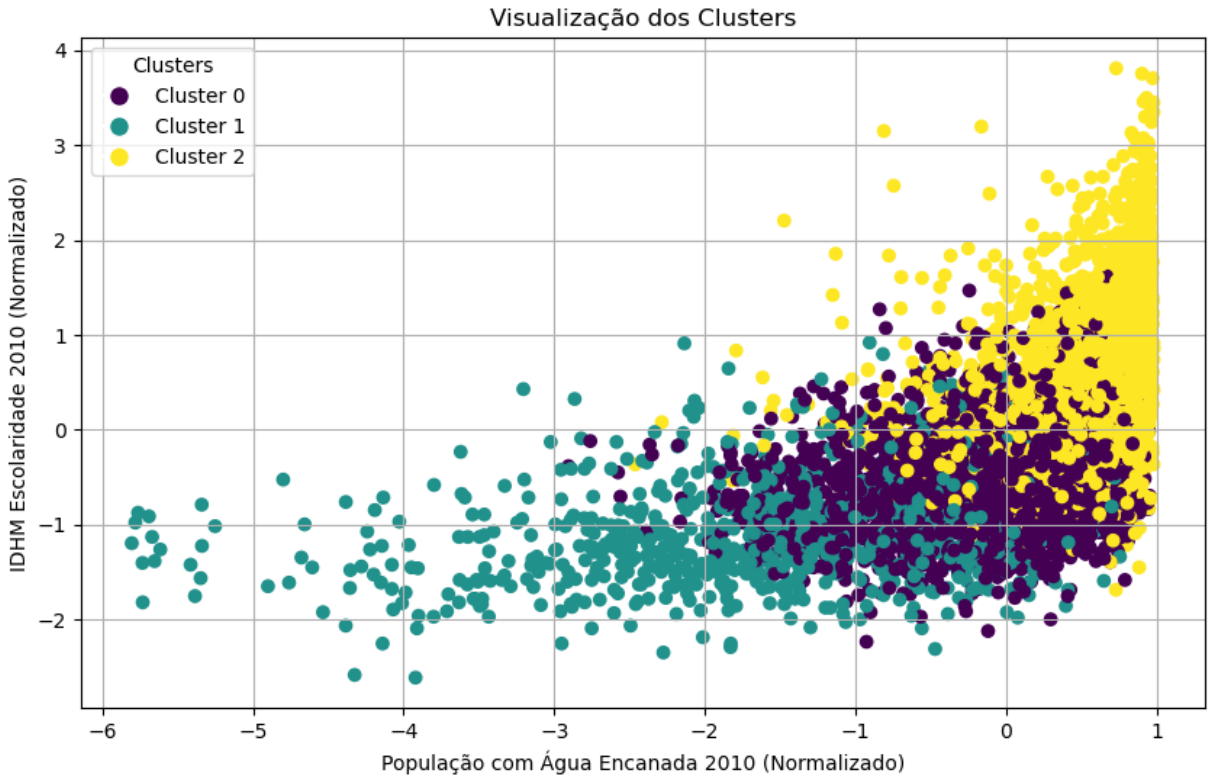

Figura 12. Distribuição dos clusters por acesso à água encanada em 2010 (normalizado) e população com mais de 18 anos com ensino fundamental completo em 2010 (normalizado)
Fonte: Elaborado pelo autor

A Figura 13 traz o gráfico de pares entre as variáveis, o qual revela uma forte associação entre saneamento e educação. Esse resultado evidencia que os municípios com melhor infraestrutura de saneamento tendem a apresentar também melhores indicadores educacionais, reforçando a interdependência entre o acesso a serviços básicos e o desempenho escolar sendo possível identificar a distribuição dos três grupos de municípios.

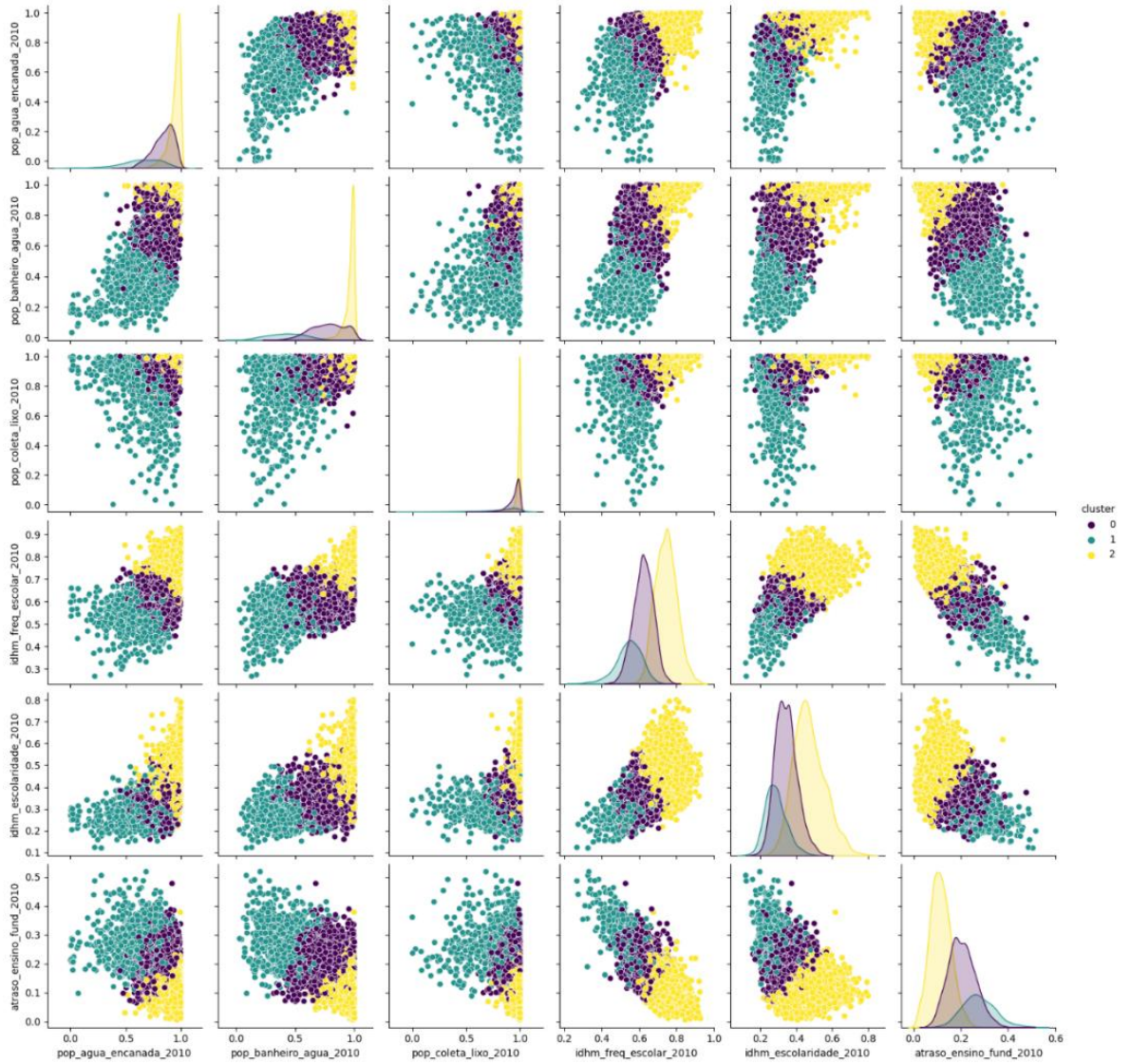

Figura 13. Gráfico de pares entre as variáveis analisadas
Fonte: Elaborado pelo autor

O Cluster 0, composto por municípios em situação intermediária, apresenta uma maior variabilidade nas suas relações. Isso indica que alguns municípios conseguem alcançar bons resultados educacionais, mesmo com um acesso limitado a saneamento, sugerindo possíveis diferenças em outros fatores que influenciam a qualidade da educação.

Ao aplicar a matriz de correlação de Pearson das variáveis nos clusters, é possível verificar relações lineares dentro de cada cluster, com isso pode-se identificar como as variáveis estão correlacionadas dentro de cada agrupamento.

Dentro do Cluster 0, conforme a Figura 14, pode-se observar uma correlação baixa, porém positiva entre a população com acesso a água encanada e banheiro com coleta de lixo. Destaca-se a correlação negativa entre frequência escolar e atraso no ensino

fundamental, mostrando que a frequência escolar é um fator relevante para evitar reprovação e atraso escolar.

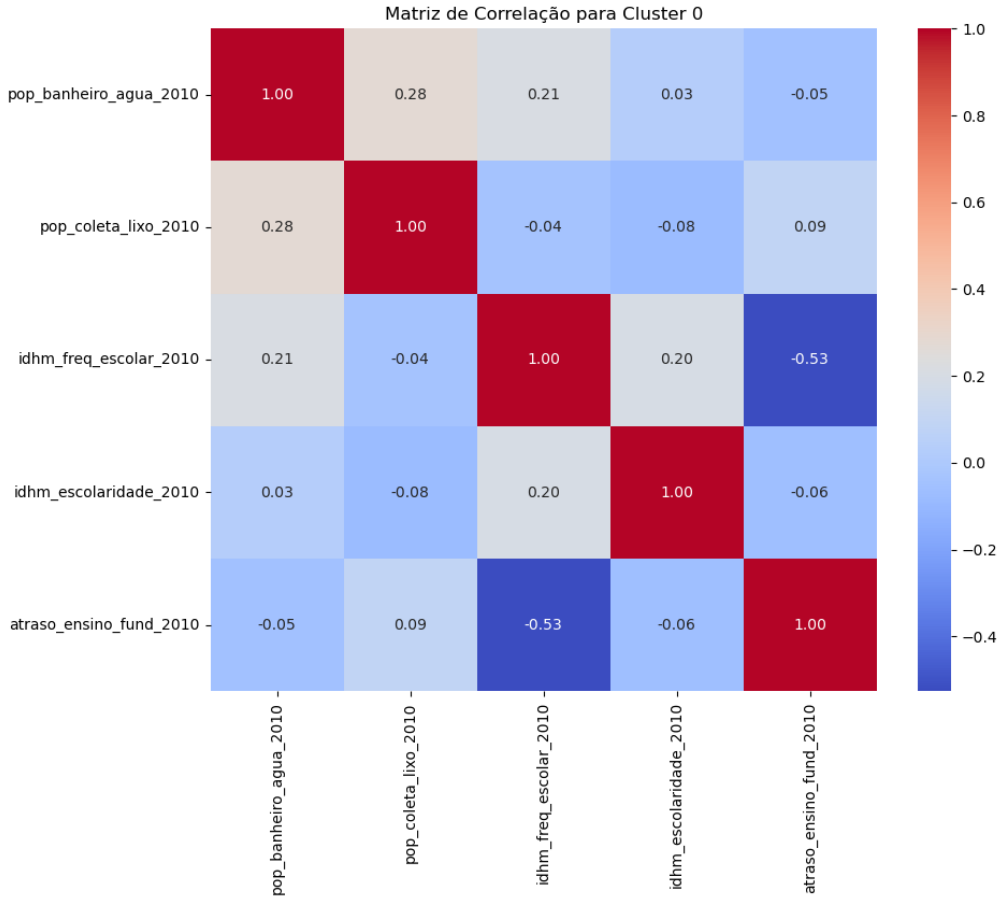

Figura 14. Matriz de correlação para o Cluster 0
Fonte: Elaborado pelo autor

No Cluster 1, conforme a Figura 15, observa-se uma correlação negativa ainda maior do que no Cluster 0 entre a frequência escolar e o atraso no ensino fundamental, mais uma vez apontando que a frequência escolar desempenha um papel importante para evitar o atraso escolar.

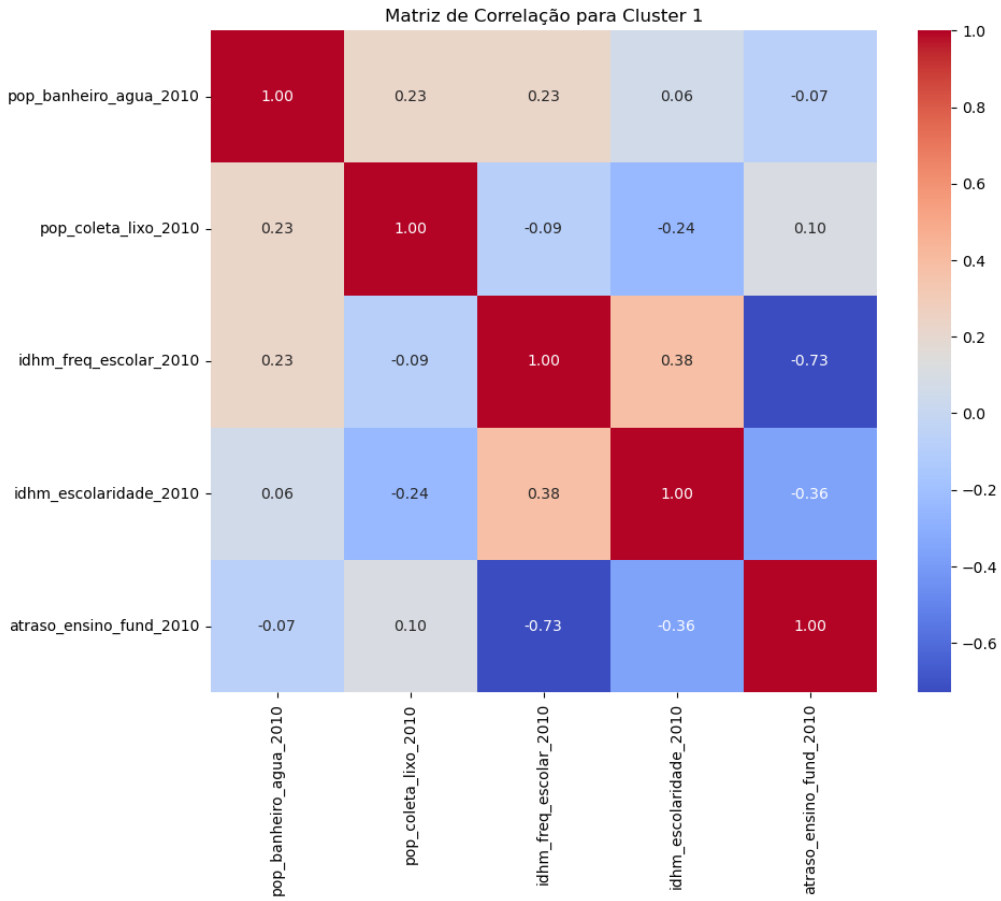

Figura 15. Matriz de Correlação para o Cluster 1
Fonte: Elaborado pelo autor

Dentro do Cluster 2, conforme a Figura 16, observou-se novamente a alta correlação negativa entre frequência escolar e atraso no ensino fundamental. Também ocorre uma maior correlação entre a coleta de lixo e o acesso a água encanada e banheiro na residência. Porém o que destoou mais no Cluster 2 em relação aos outros é a presença de correlação positiva entre o acesso a coleta de lixo e a escolaridade.

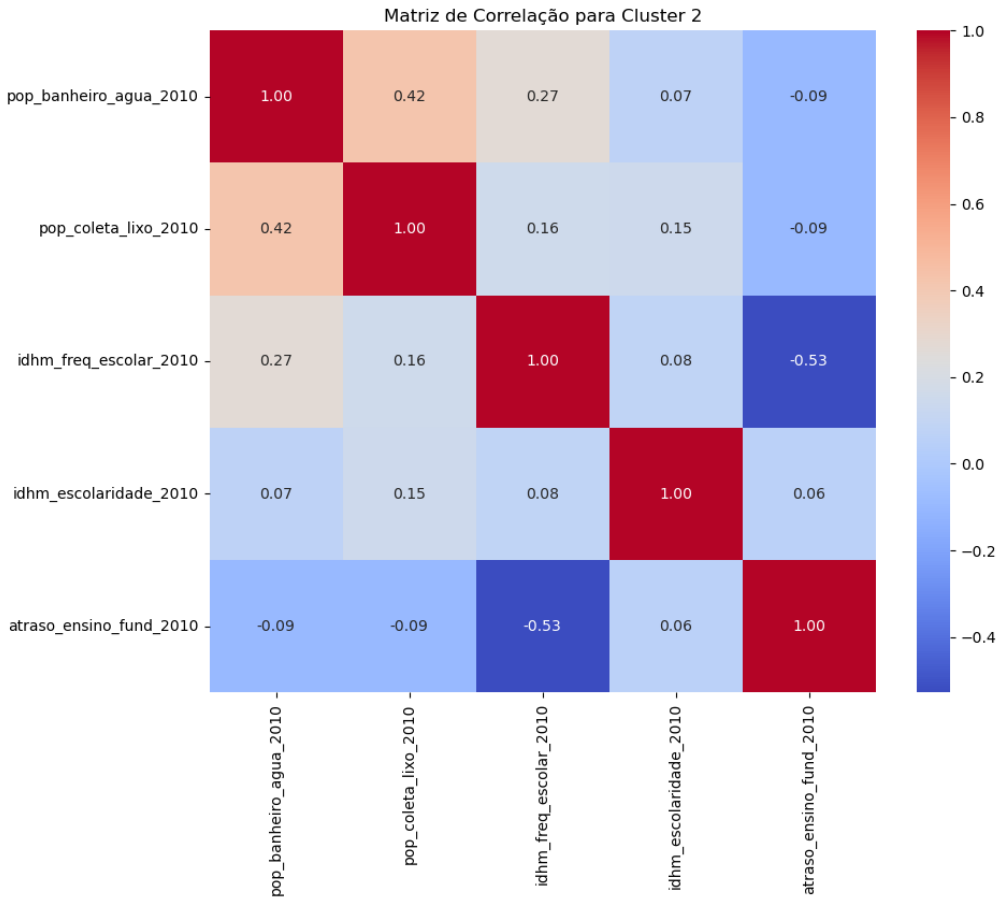

Figura 16. Matriz de Correlação para o Cluster 2
Fonte: Elaborado pelo autor

Com o acesso a uma melhor condição de saneamento estando associado positivamente à escolaridade, foi aplicada uma variável de renda nos parâmetros para verificar o comportamento dos clusters. Para evitar distorções na renda média pela porcentagem mais rica da população, optou-se pela renda do quarto quinto mais pobre da população, conforme apresentado na Figura 17.

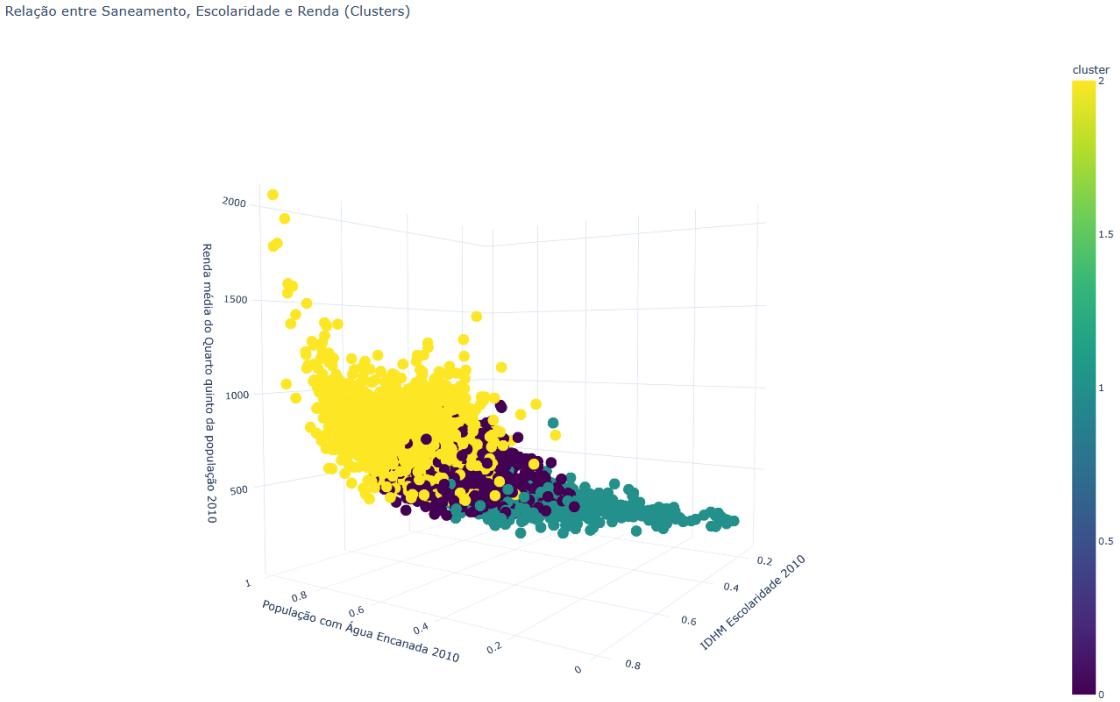

Figura 17. Distribuição dos Clusters por acesso à água encanada (2010), população com mais de 18 anos e ensino fundamental completo (2010) e renda média do quarto quinto mais pobre da população (2010)
Fonte: Elaborado pelo autor

Foi observado que o comportamento geral dos clusters se manteve, com o fator renda mantendo o padrão observado nos municípios, isto é, municípios onde a renda média é maior, possuem um maior acesso a saneamento básico e maior escolaridade para a população.

Os resultados obtidos confirmam a hipótese inicial de que há uma correlação positiva entre renda, saneamento básico e escolaridade. Municípios com maior renda média tendem a apresentar melhores condições de acesso à água tratada, coleta de lixo e índices educacionais mais elevados. O comportamento dos clusters reforça que, em locais com menor renda, a infraestrutura de saneamento é mais precária, o que impacta diretamente os indicadores educacionais. A análise também sugere que as disparidades regionais no Brasil, historicamente associadas à renda, permanecem como um fator preponderante para a distribuição desigual de recursos e oportunidades.

A utilização de técnicas de análise de clusters possibilitou a segmentação dos municípios de maneira mais eficaz, permitindo identificar padrões subjacentes e grupos homogêneos em relação ao saneamento e à educação. Isso é particularmente útil para a criação de políticas públicas direcionadas, já que os grupos com maiores deficiências podem ser facilmente reconhecidos. Outro benefício é a capacidade de entender como múltiplas variáveis (renda, escolaridade, saneamento) interagem, fornecendo insights importantes sobre onde esforços conjuntos podem ser mais eficazes.

Embora o modelo tenha se mostrado eficiente para identificar padrões e correlações, ele apresenta algumas limitações. Primeiro, a dependência dos dados do Censo de 2010, que pode não refletir as mudanças recentes nos municípios. Além disso, a escolha de variáveis, como a renda do quarto quinto mais pobre, foi feita para evitar distorções, mas outras variáveis socioeconômicas relevantes podem ter sido desconsideradas. Por fim, a análise de clusters, embora poderosa, não capta todas as nuances das desigualdades regionais, já que não leva em consideração fatores culturais ou históricos que também podem influenciar o desenvolvimento local.

Conclusão

A análise de clusters aplicada aos municípios brasileiros revelou padrões significativos de desigualdade no acesso a saneamento e educação, refletindo as disparidades regionais do país. Os resultados demonstraram que o Cluster 2 agrupa os municípios com os maiores índices de desenvolvimento, caracterizados por elevados níveis de acesso a água encanada e escolaridade. Em contrapartida, o Cluster 1 inclui os municípios mais vulneráveis, apresentando baixos indicadores de saneamento e educação, enquanto o Cluster 0 abrange municípios em condições intermediárias, com uma maior variabilidade nos indicadores analisados.

Foi evidenciada uma correlação positiva entre saneamento e educação, reforçando a interdependência entre esses fatores. Nos municípios mais desenvolvidos, observou-se que o acesso a serviços básicos, como coleta de lixo e água encanada, está associado a melhores indicadores educacionais. A inclusão da variável de renda no modelo confirmou que municípios com maior renda média tendem a apresentar melhores condições de saneamento e níveis educacionais mais elevados, demonstrando a influência de fatores econômicos no desenvolvimento dessas áreas.

Os achados indicam a necessidade de políticas públicas que promovam o desenvolvimento integrado entre infraestrutura básica e educação, de forma a reduzir as desigualdades regionais. Municípios que pertencentes ao Cluster 1 demandam intervenções específicas voltadas à melhoria do saneamento básico e ao aumento da frequência escolar, com o objetivo de diminuir as disparidades observadas e promover uma distribuição equitativa de oportunidades.

Em suma, a análise sugere que a melhoria da infraestrutura de saneamento nos municípios brasileiros tem impacto direto nos indicadores educacionais, sendo um fator crucial para o desenvolvimento social e econômico dessas localidades. Futuras pesquisas podem aprofundar essa relação por meio de uma análise temporal, incorporando dados mais recentes, como o Censo de 2022, para observar a evolução do saneamento e da educação ao longo do tempo. Além disso, a integração de variáveis socioeconômicas adicionais, como saúde e acesso à tecnologia, pode fornecer uma visão mais completa das desigualdades regionais. Estudos que utilizem métodos mais avançados de machine learning, como redes neurais, podem refinar os agrupamentos e revelar padrões mais complexos. Por fim, a análise do impacto de políticas públicas e a utilização de simulações preditivas podem gerar insights valiosos para a formulação de intervenções governamentais mais eficazes e direcionadas.

Agradecimentos

Agradeço a Deus pela orientação e força em minha jornada, à minha mãe pelo amor e apoio incondicional, e ao meu orientador, Miguel Ângelo Lellis Moreira, pela valiosa orientação e incentivo ao longo deste trabalho.

Referências

Abanyie, S. K., Amuah, E. E. Y., Douti, N. B., Owusu, G., Amadu, C. C., et al. 2021. *WASH in Selected Basic Schools and Possible Implications on Health and Academics: An Example of the Wa Municipality of Ghana, West Africa*. American Journal of Environmental Science and Engineering, 5(1): 15-20. DOI: 10.11648/j.ajese.20210501.13.

Aggarwal, C. C.; Reddy, C. K. 2013. *Data Clustering: Algorithms and Applications*. 1. ed. Boca Raton: CRC Press.

Atlas do Desenvolvimento Humano no Brasil. 2022. *Elaboração: Atlas do Desenvolvimento Humano no Brasil*. PNUD Brasil, IPEA e FJP. Fontes: dados do IBGE e de registros administrativos. Disponível em: http://atlasbrasil.org.br/acervo/biblioteca. Acesso em: 07 set. 2024.

Bação, P., Mazeron, D., & Simões, M. 2023. *The financialization of health and education and inequality in twenty-first century Brazil*. Latin American Perspectives, 50(5): 47-66.

Brasil. 1988. *Constituição da República Federativa do Brasil. Brasília*, DF: Presidência da República. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 16 set. 2024.

Brasil. 2016. *Proposta de Emenda à Constituição nº 2, de 2016*. Altera o art. 206 da Constituição Federal. Brasília. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/124779. Acesso em: 16 set. 2024.

Brasil. 2020. *Lei nº 14.026, de 15 de julho de 2020*. Disponível em: https://normas.leg.br/?urn=urn:lex:br:federal:lei:2020-07-15;14026. Acesso em: 10 jun. 2024.

Bruce, A.; Bruce, P. 2019. *Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais*. Alta Books, Rio de Janeiro, RJ, Brasil.

Cournapeau, D. 2010. *Scikit-learn*. Disponível em: https://scikit-learn.org/stable/. Acesso em: 10 jun. 2024.

Fávero, L. P.; Belfiore, P. 2024. *Manual de Análise de Dados*. 2ª edição. GEN LTC, Brasil.

Géron, A. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., Sebastopol, CA, EUA.

Hunter, J. D. 2002. *Matplotlib*. Disponível em: https://matplotlib.org/. Acesso em: 10 jun. 2024.

IBGE. Coordenação de Estruturas Territoriais. *Malha Municipal Digital e Áreas Territoriais 2023: notas metodológicas 01/2024*. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?=&t=acesso-ao-produto. Acesso em: 16 set. 2024.

Instituto Trata Brasil. 2023. *Guia do Saneamento 2023*. São Paulo: [s.n.]. Disponível em: https://tratabrasil.org.br/wp-content/uploads/2024/04/Guia-do-Saneamento-2023_V20_12.11_Digital.pdf. Acesso em: 16 set. 2024.

Larson, R.; Farber, B. 2015. *Estatística Aplicada*. Tradução de José Fernando Pereira Gonçalves; revisão técnica de Manoel Henrique Salgado. Pearson Education do Brasil, São Paulo, SP, Brasil.

McKinney, W. W. 2010. *Pandas*. Disponível em: https://pandas.pydata.org/. Acesso em: 10 jun. 2024.

McKinney, W. W. 2018. *Python para Análise de Dados*. 2ª edição. Pearson Education do Brasil, São Paulo, SP, Brasil.

Montgomery, D. C.; Runger, G. C. 2014. *Estatística Aplicada e Probabilidade para Engenheiros*. 6. ed. Wiley.

Nainggolan, R., et al. 2019. *Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method*. Journal of Physics: Conference Series, 1361: 012015. IOP Publishing. DOI: 10.1088/1742-6596/1361/1/012015.

OECD. 2020. How's Life? 2020: Measuring Well-being. Paris: OECD Publishing. Disponível em: https://doi.org/10.1787/9870c393-en. Acesso em: 16 set. 2024.

Prakash, K. B. 2022. *Data Science Handbook: A Practical Approach*. John Wiley & Sons. DOI: 10.1002/9781119858010.

Reitz, K.; Schlusser, T. 2017. *O Guia do Mochileiro Python: Melhores Práticas Para Desenvolvimento*. Novatec Editora, Brasil.

Sharma, M. K.; Adhikari, R.; Khanal, S. P.; Acharya, D.; Teijlingen, E. v. 2024. *Do school water, sanitation, and hygiene facilities affect students' health status, attendance, and educational achievements?* Health Science Reports, 7: e2293. DOI: 10.1002/hsr2.2293.

Valencio, N. A., & Baptista, M. S. 2023. *The Interface of Disasters, Sanitation, and Poverty in Brazil: A Sociological Perspective*. Frontiers in Sustainable Cities, 5: 1184532. DOI: 10.3389/frsc.2023.1184532.

Waskom, M. 2012. *Seaborn*. Disponível em: https://seaborn.pydata.org/. Acesso em: 10 jun. 2024.

Apêndice

LIMA, Pablo H. S. TCC-MBA-USP-ESALQ-DSA-231. Disponível em: https://github.com/limapablo/TCC-MBA-USP-ESALQ-DSA-231. Acesso em: 16 set. 2024.