Análise da desigualdade socioeconômica da cidade de São Paulo, a partir de técnicas multivariadas exploratórias, através de métodos de PCA e Análise de Cluster

Ernesto Gondim Neiva[1]; Rodrigo Bruno Zanin[2]

[1] Aluno do Curso de MBA Data Science e Analytics na Escola Superior de Agricultura "Luiz de Queiroz, Universidade de São Paulo". Avenida Pádua Dias, 235, Agronomia, CEP 13418-900, Piracicaba, São Paulo, Brasil.
[2] Universidade do Estado de Mato Grosso - UNEMAT. Professor Doutor da Faculdade de Ciências Exatas e Tecnológicas, Campus de Sinop/MT. Rua T, Número 01, Jardim Santa Cruz II, 78077-033, Cuiabá, Mato Grosso, Brasil.

Análise da desigualdade socioeconômica da cidade de São Paulo, a partir de técnicas multivariadas exploratórias, através de métodos de PCA e Análise de Cluster

Resumo

A cidade de São Paulo é conhecida por seus contrastes entre bairros ricos e pobres, e a desigualdade é uma questão crônica que afeta a vida de milhões de pessoas. É fundamental destacar que a desigualdade também apresenta aspectos raciais, já que a população preta é historicamente desfavorecida e continua enfrentando barreiras para o acesso a bens e serviços básicos. Obter dados para compreender a magnitude e a natureza desse problema é fundamental. Porém, lidar com um grande volume de informações não é simples, já que os dados são complexos e difíceis de interpretar. Além disso, a falta de padronização na coleta e na apresentação de dados pode tornar a análise ainda mais desafiadora. Nesse contexto, as técnicas de estatística não supervisionadas, como PCA (Análise de Componentes Principais) e Análise de Cluster, podem ser utilizadas para realizar tarefas como a redução de dimensionalidade, a identificação de padrões e tendências nos dados, a visualização de grupos ou clusters de observações similares. O objetivo é mostrar como essas técnicas podem ser aplicadas para analisar dados sociais e econômicos e identificar padrões e tendências que possam ser utilizados para compreender e abordar a desigualdade na cidade. Por fim, conclui-se que a combinação de análise de dados e técnicas estatísticas pode ser uma ferramenta eficaz para ajudar a lidar com desafios sociais complexos, auxiliando na criação de políticas públicas no combate à desigualdade socioeconômica.

Palavras-chave: Aprendizagem não-supervisionada; Análise por componentes principais; Análise de agrupamentos; IDH; Racismo.

Abstract

The city of São Paulo is known for its contrasts between rich and poor districts,where inequality is a chronic issue affecting the lives of millions of people. It is crucial to highlight that inequality also has racial aspects, as the black population is historically disadvantaged and continues to face barriers to access to basic goods and services. Obtaining data to understand the magnitude and nature of this problem is essential. However, dealing with a large amount of information is not simple, as data can be complex and difficult to understand. Moreover, the lack of standardization in data collection and presentation can make analysis even more challenging. In this context, unsupervised statistical techniques such as PCA (Principal Component Analysis) and Cluster Analysis can be used to perform tasks such as dimensionality reduction, identification of patterns and trends in data, visualization of groups or clusters pairing up similar observations. The purpose is to show how these techniques can be applied to analyze social and economic data and identify patterns and trends that can be used to understand and address inequality in the city. In conclusion, the combination of data analysis and statistical techniques can be an effective tool to help deal with complex social challenges, assisting in the creation of public policies to fight socioeconomic inequality.

Keywords: Unsupervised learning; Principal component analysis; Cluster analysis; HDI; Racism.

1 - **Introdução**

A desigualdade social é um problema complexo e multifacetado. Nas palavras de Florestan Fernandes, ela é “um obstáculo para o desenvolvimento econômico, político e social do país, e impede o acesso à justiça, à educação, à saúde e a outros direitos básicos da população." Para combatê-la, são requeridas ações e políticas públicas integradas e bem-sucedidas. Logo, são necessários instrumentos que forneçam dados necessários para a identificar as áreas de ação.

Segundo o Programa das Nações Unidas para o Desenvolvimento (PNUD), as três dimensões essenciais para o desenvolvimento humano são: “a oportunidade de viver uma vida longa e saudável, de ter acesso ao conhecimento e ter um padrão de vida que garanta as necessidades básicas, representadas pela saúde, educação e renda.”

Criado em 1990, o Índice de Desenvolvimento Humano (IDH) é a medida amplamente utilizada para verificar o progresso e o bem-estar de um país. Em comparação com o Produto Interno Bruto (PIB), cujo foco está somente no crescimento econômico, o IDH oferece uma visão mais completa dos aspectos sociais. No entanto, ele pode não refletir adequadamente as desigualdades sociais existentes em uma área.

Para além dos dados utilizados no cálculo IDH, o entendimento da qualidade de vida e sua distribuição entre os diferentes extratos da sociedade requer análise de outras características. Os dados sobre o Desenvolvimento Humano podem ser utilizados como parte de um conjunto informacional objetivo para uma investigação social mais profunda sobre desigualdade social.

De acordo com o Atlas do Desenvolvimento Humano no Brasil (2013), o "Desenvolvimento humano é o processo de ampliação das liberdades das pessoas, no que tange suas capacidades e as oportunidades a seu dispor, para que elas possam escolher a vida que desejam ter." Porém, por si só, podem não revelar o cenário real de determinadas regiões.

Uma combinação de fatores históricos, sociais e culturais, - como a escravidão o colonialismo europeu - determinou a situação material dos diferentes grupos sociais no Brasil. A escravidão deixou uma marca profunda na sociedade brasileira. O resultado dessas estruturas sociais e históricas refletem na desigualdade racial em relação ao poder econômico. Nesse sentido, Silvio Almeida (2018) explica que:

> *"(...) a raça é um marcador determinante da desigualdade econômica e que direitos sociais e políticas universais de combate à pobreza e distribuição de renda que não levem em conta o fator raça/cor mostram-se poucos efetivas."*

Assim, a compreensão do movimento real das dinâmicas sociais deve levar em conta suas raízes.

Composta por 96 distritos, a cidade de São Paulo é considerada o mais importante centro econômico do Brasil. Em contraste, é marcada por uma enorme disparidade na qualidade de vida entre as diferentes áreas da cidade. Como salienta o sociólogo Américo Sampaio, “os indicadores que deflagram as piores condições de vida na cidade e os menores acessos aos serviços e políticas públicas municipais estão sobre os territórios dos distritos mais afastados do centro e com maior proporção de população jovem, preta e parda. Esse elemento não pode ser desconsiderado dessa análise.”

O objetivo deste trabalho é determinar a relação entre as características socioeconômicas dos bairros de São Paulo. Através desta análise, espera-se compreender como a desigualdade social afeta o desenvolvimento humano e identificar as áreas da cidade que precisam de atenção especial para promover um desenvolvimento mais equilibrado. Além disso, espera-se contribuir para a discussão sobre a importância de medir e abordar as desigualdades sociais em conjunto com o desenvolvimento humano.

2 - **Material e Métodos**

O presente trabalho foi conduzido utilizando os dados do último Censo demográfico brasileiro (2010) do IBGE (Instituto Brasileiro de Geografia e Estatística). A coleta de dados foi realizada por meio de questionários preenchidos por entrevistadores treinados pelo IBGE, que visitam todos os domicílios brasileiros. Além disso, informações secundárias, como registros de órgãos públicos, são utilizadas para complementar e validar as informações coletadas através das entrevistas.

Para avaliar as condições de vida e o progresso humano em nível municipal, será utilizado o IDHM (Índice de Desenvolvimento Humano Municipal). O IDHM inclui três componentes: IDHM Longevidade, IDHM Educação e IDHM Renda. A forma detalhada do cálculo do IDHM não será objeto do presente trabalho, no entanto pode ser encontrada em *O índice de desenvolvimento humano municipal brasileiro* (2013), livro disponível no site do Ipea.

Colaborativamente, será utilizado o Mapa da Desigualdade de São Paulo. Criado pela organização da sociedade civil Rede Nossa São Paulo (RNSP), o mapa tem “a proposta de revelar a cidade por meio de suas diferenças regionais (alcançando a menor unidade administrativa possível), a ferramenta apresenta uma seleção de indicadores técnicos essenciais por distrito (...)”.

A análise de dados, geração de estatísticas e plotagem de tabelas e gráficos foi realizada com a linguagem R.

Os resultados do trabalho podem ajudar no planejamento de futuras análises. Porém, não darão suporte à criação de políticas públicas visto que a desatualização dos dados não reflete as reais condições do objeto em estudo.

A metodologia utilizada para analisar a relação entre os dados de IDH dos bairros de São Paulo e Mapa da Desigualdade seguirá as seguintes etapas:

1. Limpeza e preparação dos dados: verificação da completude, precisão e formatação adequada dos dados. Isso pode incluir a correção de erros, a eliminação de duplicatas e a normalização dos dados.
2. Análise exploratória de dados (EDA): análise das distribuições das variáveis em diferentes bairros de São Paulo, a fim de entender a amplitude e a variação dos dados. Plotagem das variáveis no mapa da cidade usando a geolocalização dos bairros para plotar os valores de IDH e as variáveis presentes no mapa da desigualdade, a fim de visualizar as relações espaciais entre as variáveis.
3. Análise de correlação: medir a força e a direção da relação linear entre os dados de IDH e as variáveis presentes no Mapa da Desigualdade. Serão utilizados gráficos e tabelas para identificar padrões e relações entre as variáveis.
4. Análise de componentes principais (PCA): redução da dimensionalidade dos dados, facilitando a visualização e a análise das relações entre os bairros.
5. Análise de Cluster: Divisão dos bairros em grupos com base na semelhança dos valores obtidos através da técnica de PCA.

2.1 - **Limpeza e preparação dos dados**

Os dados utilizados estão representados nas Mapa da Desigualdade da cidade de São Paulo e no Censo 2010 do IBGE. Inicialmente, serão excluídas as linhas com os dados do ano 2000 e mantidas apenas as de 2010.

$$dados\_bairros\_2010 < -dados\_bairros[dados\_bairros\$ANO == 2010,]$$

Em seguida, será criado um *dataset* com apenas uma variável quantitativa, o IDHM.

$$dados\_bairros\_IDHM < -dados\_bairros\_2010\% > \%select("DISTRITO","CODDIST","REGIAO8","IDHM")$$

Com as funções:

$$sum(is.na(dados\_bairros\_IDHM))$$

$$sum(is.null(dados\_bairros\_IDHM))$$

É possível verificar a existência de observações vazias e nulas. Duas linhas vazias foram identificadas e removidas com a função:

$$dados\_bairros\_IDHM < -na.omit(dados\_bairros\_IDHM)\,.$$

	DISTRITO	CODDIST	REGIAO8	IDHM
1595	Vila Formosa	85	Leste 1	0.866
1596	Carrão	20	Leste 1	0.870
1597	Carrão	20	Leste 1	0.790
1598	Vila Formosa	85	Leste 1	0.816
1599	Aricanduva	4	Leste 1	0.820
1600	Aricanduva	4	Leste 1	0.870

Tabela 1:Parte do Dataset com dados de 2010 e IDHM dos bairros de São Paulo

Cada bairro de São Paulo possui diferentes UDHs (Unidades de Desenvolvimento Humano). As UDHs apresentam uma visão geral dos locais dentro dos municípios brasileiros, reunindo informações comparativas em um único conjunto de dados. Para reunir as informações em uma única observação por bairro será feita a média dos valores de IDHM nos bairros repetidos:

$$IDHMBairro = \frac{0.870 + 0.790(Carrão)}{2} + \frac{0.820 + 0.870(Aricanduva)}{2} + \cdots$$

Para incluir as informações do Mapa da Desigualdade, suas variáveis serão renomeadas para a criação de um *dataset* único com a função *full_join:*

$$dados\_bairros\_completo <- full\_join(dados\_bairros\_unicos, mapa\_desigualdade)$$

Em seguida, os valores nas colunas de interesse devem ser transformados em numéricos

$$lapply(dados\_bairros\_completo[, -c(1:3)], function(x) as.numeric(as.character(x)))$$

E padronizados com o *Z-score*. Conforme descrito por Fávero e Belfiore (2017), “o processo de padronização *Zscores* fará com que todas as novas variáveis padronizadas apresentem média zero e variância 1”. Com este procedimento, os dados passam a ser uma distribuição padrão, o que pode ser útil em muitas aplicações, como análise de dados, comparação entre amostras e identificação de *outliers* (Figura 2). O cálculo é feito da seguinte forma:

$$Z = \frac{(x - \mu)}{\sigma}$$

Onde *x* é o valor da observação, *μ* representa a média e *σ* é o desvio padrão. A variável original é transformada em uma nova variável aleatória *Z,* com média zero (μ = 0) e variância

1 ($a^2 = 1$).

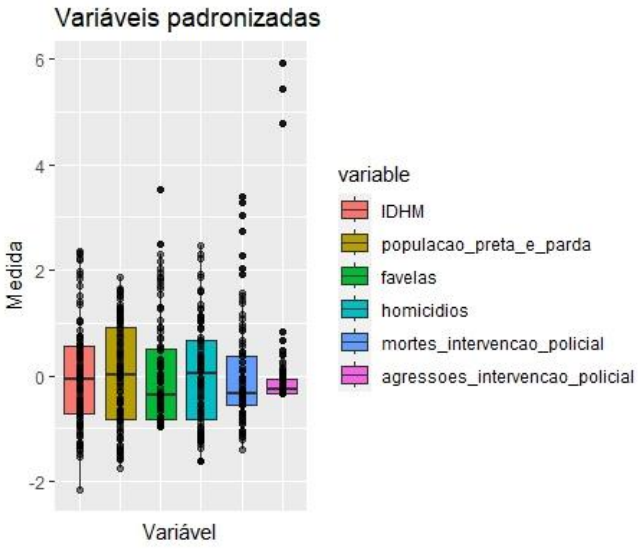

Figura 1: Variáveis padronizadas

A plotagem dos dados no mapa da cidade de São Paulo ajudará a compreender as relações espaciais entre as variáveis e as tendências geográficas de forma mais clara. O mapa pode ser usado para explorar padrões e relações não óbvias nos dados e para identificar pontos de interesse para análises adicionais.

A biblioteca *geobr* é usada para acessar dados geográficos do Brasil no R. Ela oferece uma interface para acessar dados de fontes oficiais, como o Instituto Brasileiro de Geografia e Estatística (IBGE) e o Instituto Nacional de Meteorologia (INMET), e permite a manipulação e análise destes dados de forma fácil e intuitiva. Além disso, a *geobr* fornece funções para plotar mapas temáticos e visualizar dados geográficos de forma gráfica. Segundo o texto retirado do site do github do pacote (https://github.com/ipeaGIT/geobr), “pacote inclui uma ampla gama de dados geoespaciais em formato de geopacote (como *shapefiles*), disponíveis em várias escalas geográficas e por vários anos com atributos harmonizados, projeção e topologia.”

A criação de um *shapefile* com a localização geográfica somente dos bairros da cidade de São Paulo é feita através do carregamento do mapa do Brasil (1), filtragem dos dados da cidade (2) e a combinação do *dataset* preparado previamente com as informações geográficas do *shapefile* (3).

$$maps\_bairros <- geobr::read\_neighborhood() \quad (1)$$

$$maps\_bairros <- maps\_bairros \%>\% filter(str\_detect(name\_muni, "S.oPaulo\$")) \quad (2)$$

$$dados\_bairros\_mapeados <- maps\_bairros \%>\% full\_join(dados\_bairros\_completo) \%>\% select\_if(\sim !any(is.na(.))) \quad (3)$$

2.2 - **Análise exploratória dos dados**

A análise exploratória de dados (AED) é uma etapa importante no processo de análise de dados. Ela tem como objetivo compreender o conjunto de dados, identificar tendências, padrões, outliers e possíveis problemas ou irregularidades nos dados. Além disso, a AED é uma oportunidade para se familiarizar com os dados e formular hipóteses sobre como as variáveis estão relacionadas entre si.

Mapas com indicadores sociais e econômicos permitem comparações entre diferentes localidades, sendo importante para identificar áreas que precisam de maior atenção ou apoio. A seguir está apresentada a distribuição de cada variável do conjunto de dados no mapa da cidade de São Paulo. Para a plotagem no formato *shapefile* é utilizado a função *ggplot.*

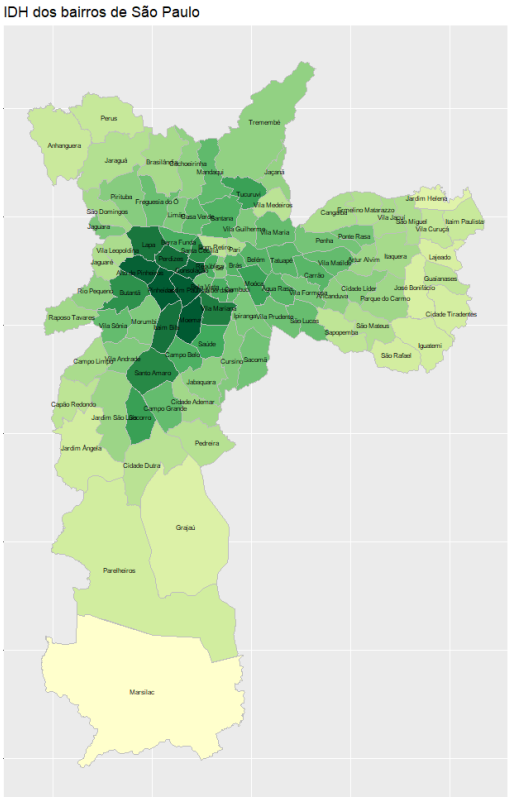

Figura 2: IDH

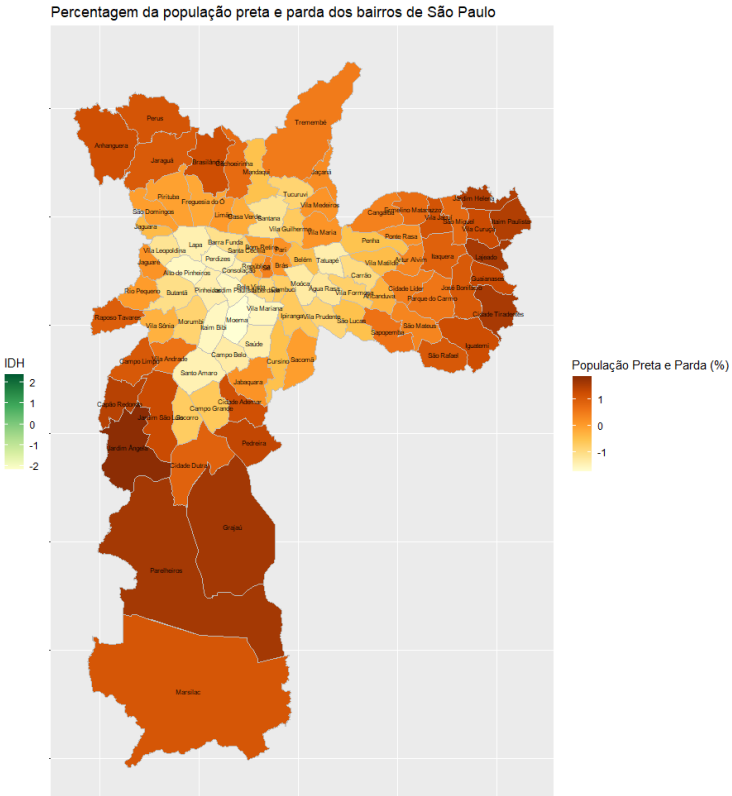

Figura 3: Percentagem de opulação preta e parda

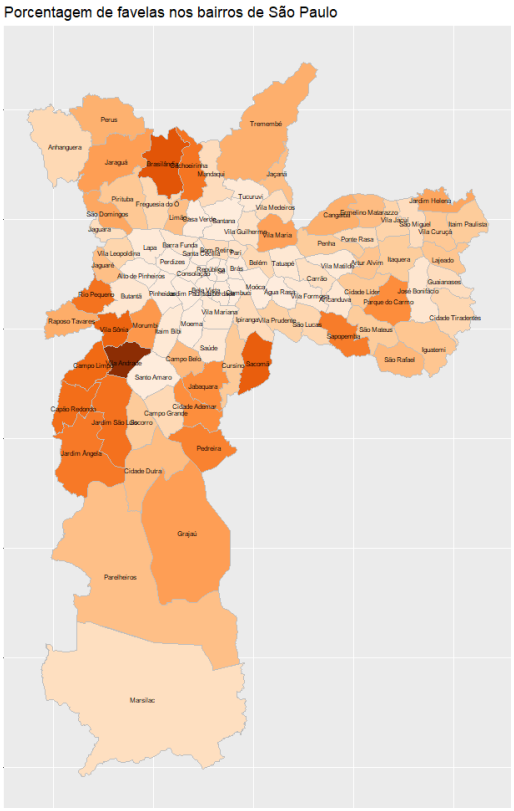

Figura 4: Favelas

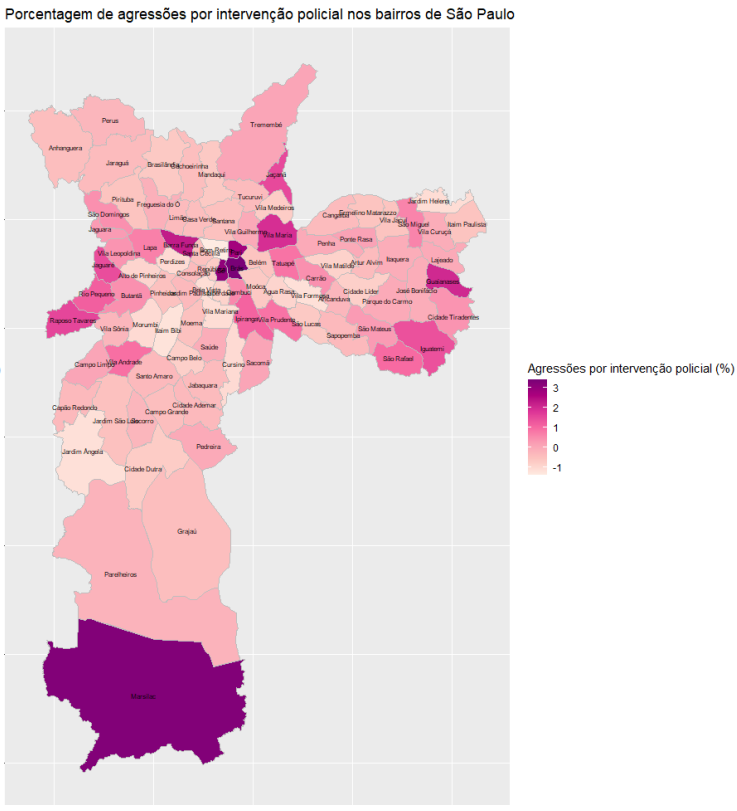

Figura 5: Agressões por intervenção policial

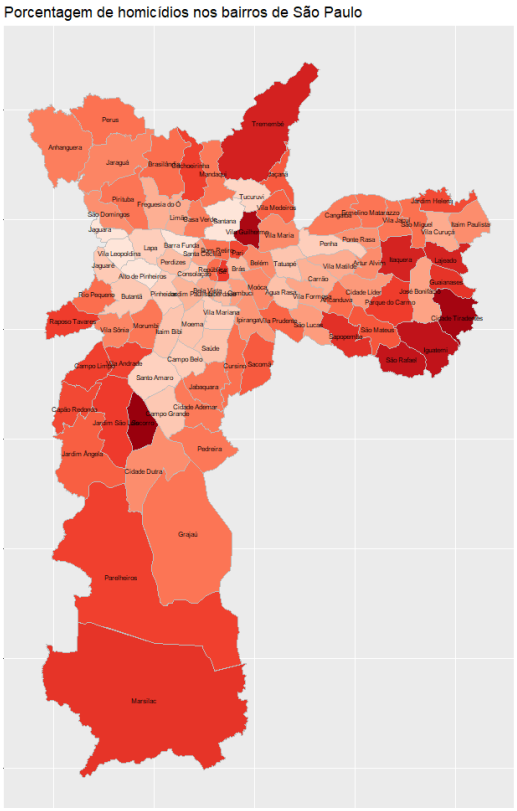

Figura 6: Homicídios

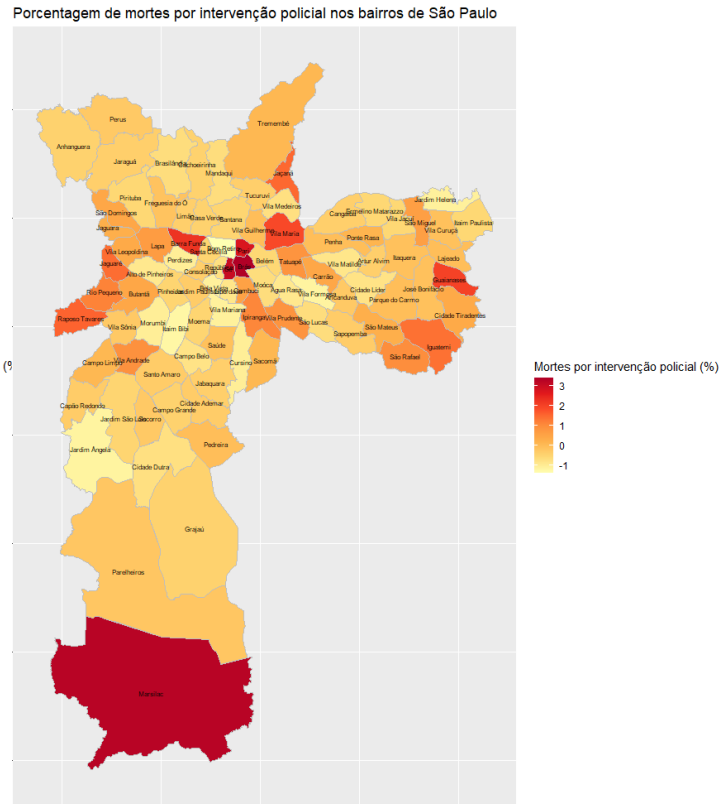

Figura 7: Mortes por intervenção policial

2.3 - **Análise de correlação**

Correlação é uma medida estatística que descreve a relação entre duas variáveis quantitativas. A função *cor* no R é uma função que pode ser usada para calcular a matriz de correlação entre duas ou mais variáveis. Ela usa o método de Pearson para calcular o coeficiente de correlação linear entre cada par de variáveis, que varia de -1 a 1 e indica o grau de relação linear entre as variáveis. Quando o coeficiente de Pearson é próximo de 1, há uma forte correlação positiva linear entre as variáveis. Quando o coeficiente é próximo de -1, há uma forte correlação negativa linear. Um coeficiente de correlação de 0 indica ausência de correlação. De acordo com Fávero e Belfiore (2017), o coeficiente de correlação de Pearson *(p)* pode ser calculado como a razão entre a covariância de duas variáveis e o produto dos desvios-padrão *(S)* de cada uma delas, conforme segue:

$$\rho = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right).\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Para avaliar a significância estatística das correlações entre as variáveis, é utilizada a medida p-valor. Se o p-valor é menor que um determinado nível de significância, pode-se rejeitar a hipótese nula de independência entre as variáveis e concluir que há uma associação estatisticamente significante entre elas. “Assim, para um nível de confiança de 95%, se *P-value* < 0,05, a hipótese é rejeitada e podemos afirmar que há associação entre as variáveis” (Fávero e Belfore, 2017). Assim, utilizando uma significância de 5%, é criada uma tabela com o grau de correlação entre as variáveis (Figura 9).

Analisando o resultado dos p-valores (Tabela 2), verifica-se que não há associação significante entre *agressões por intervenção policial* e as demais variáveis. Similarmente, a variável *mortes por intervenção militar* apresenta correlação significante apenas com *IDHM*.

	IDHM	populacao_preta_e_parda	favelas	homicidios	mortes_intervencao_policial	agressoes_intervencao_policial
IDHM	0.0000000	0.0000000	0.0000103	0.0000000	0.0539856	0.9027759
populacao_preta_e_parda	0.0000000	0.0000000	0.0000000	0.0000000	0.1240671	0.9920670
favelas	0.0000103	0.0000000	0.0000000	0.0000078	0.7762697	0.1087942
homicidios	0.0000000	0.0000000	0.0000078	0.0000000	0.0562841	0.4525047
mortes_intervencao_policial	0.0539856	0.1240671	0.7762697	0.0562841	0.0000000	0.1535802
agressoes_intervencao_policial	0.9027759	0.9920670	0.1087942	0.4525047	0.1535802	0.0000000

Tabela 2: Relação do p-valor entre as variáveis

De outro modo, o *IDHM* possui um elevado grau de correlação negativa com *população preta e parda* e *homicídios,* chegando a -0.88 e -0.66, respectivamente. Ainda, há indícios de associações positivas entre *favelas* e *população preta e parda* (0.55), *favelas* e *homicídio* (0.44) e, por fim, *população preta e parda* e *homicídios* (0.66).

Figura 8: Correlação entre variáveis

Para a próxima etapa, serão desconsideradas as correlações não significativas.

2.4 - **Análise de componentes principais (PCA)**

PCA é uma técnica de análise fatorial que visa transformar uma grande quantidade de variáveis correlacionadas em um número menor de variáveis não correlacionadas (componentes principais), perdendo a menor quantidade de informação possível. De acordo com Fávero e Belfore (2017), em um contexto socioeconômico, a análise fatorial é utilizada para a redução estrutural dos dados e para posterior elaboração de um indicador socioeconómico que capte o comportamento conjunto dessas variáveis. Isso permite uma representação gráfica mais clara e interpretável dos dados, além de tornar mais fácil a identificação de padrões e relações entre as variáveis.

A aplicabilidade da PCA é condicionada ao Teste de Esfericidade de *Bartlett*, função $cortest.bartlett$ no r, um teste de hipóteses estatístico que verifica se uma matriz de covariância é igual à matriz identidade. Se a hipótese nula de que a matriz de covariância é igual à matriz identidade é rejeitada, as variáveis estão correlacionadas e a análise pode ser aplicada aos dados. Após o teste nas variáveis do *dataset*, foi obtido o p-valor de 0. Assim, a PCA pode ser empregada. Para isso, foi utilizada a função PCA. Após, é feita a análise dos autovalores e proporção de variância dos componentes gerados (Tabela 3 e Figura 8). Onde o autovalor é usado para medir a importância de cada componente principal na representação dos dados, a proporção de variância representa a porcentagem da variância total dos dados que é explicada por cada componente, - sua soma deve ser igual a 1. De acordo com a tabela, o primeiro componente representa mais de 70% da variância dos dados, enquanto a o segundo explica cerca de 15%. Os componentes 3 e 4 apresentam as proporções próximas 11% e 3%, respectivamente.

	Autovalores	Prop. da Variância	Prop. da Variância Acumulada
comp 1	2.8077273	70.193183	70.19318
comp 2	0.6357510	15.893775	86.08696
comp 3	0.4449947	11.124868	97.21183
comp 4	0.1115270	2.788174	100.00000

Tabela 3: Autovalores e proporção de variância

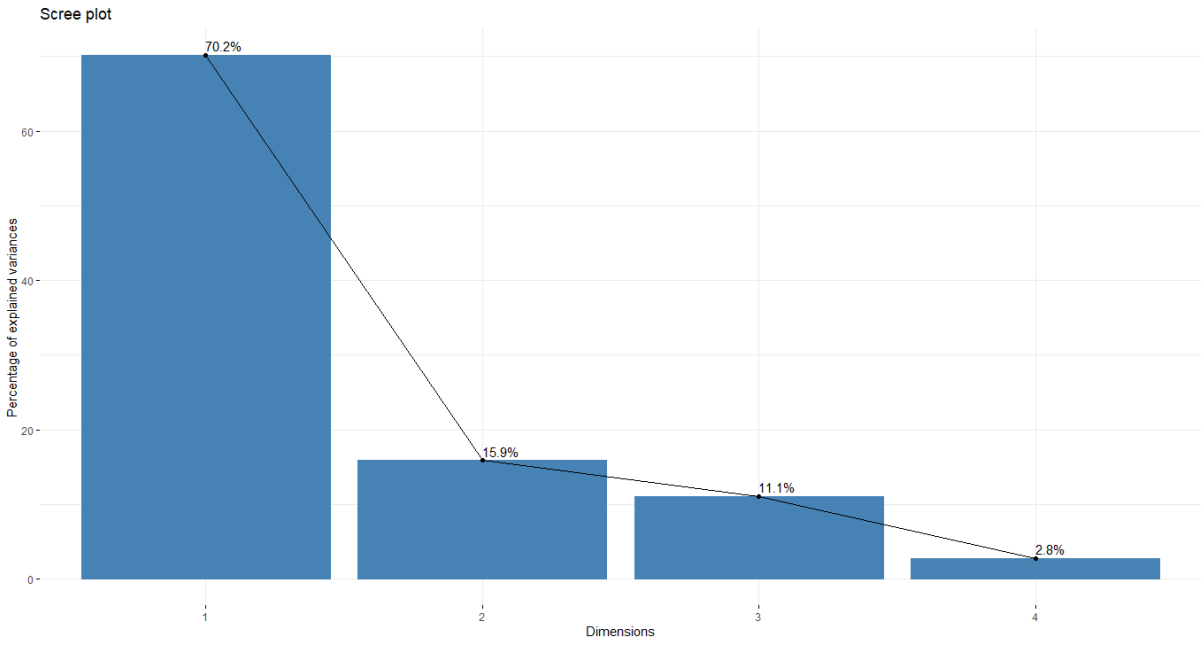

Figura 9: Proporção de variância por componente principal

A decisão sobre a quantidade de componentes a serem escolhidos foi amparada na contribuição de cada variável na formação das dimensões. Foi utilizada a função $analise\_pca\$var\$contrib$ para gerar os dados (Tabela 4).

O critério de Kaiser-Meyer-Olkin (KMO) é uma ferramenta estatística usada para determinar o número ideal de componentes principais em um conjunto de dados. Ele é baseado no valor da raiz quadrada da soma dos quadrados das autovalores. Como explica Fávero e Belfore (2017), os valores mais próximos de 1 indicam que as variáveis compartilham um percentual de variância bastante elevado (correlações de Pearson altas), valores mais próximos de 0 são decorrentes de correlações de Pearson baixas entre as variáveis, o que pode indicar que a análise fatorial será inadequada.

Apesar do critério de Kaiser sugerir a escolha de dimensões com autovalores maiores que 1, verifica-se que a variável "Favelas" possui maior representatividade no segundo componente (PC2). Assim, utiliza-se as duas primeiras dimensões para análise.

Influência de cada variável por componente

	PC1	PC2	PC3	PC4
IDHM	-0.5313931	0.3551899	-0.4276552	-0.6391967
populacao_preta_e_parda	0.5596307	-0.1746881	0.2900906	-0.7564027
favelas	0.4140216	0.9007123	0.0551159	0.1194392
homicidios	0.4827224	-0.1790015	-0.8543540	0.0708296

Tabela 4: Influência de cada variável por componente

Para visualizar a distribuição das variáveis nas duas dimensões selecionadas, foi plotado o gráfico por componentes principais com a função $fviz\_pca\_var$.

A partir da análise do gráfico (Figura 10), é possível verificar que as variáveis *IDHM*, *populacao_preta_e_parda* e *homicidios* são melhores explicadas pela Dimensão 1. Enquanto que a variável *favelas* é descrita com mais ênfase pela Dimensão 2. O resultado do gráfico corrobora visualmente com a análise de correlação vista anteriormente. Há uma relação positiva clara entre *populacao_preta_e_parda* e *homicidios*. Em contraste, na medida em que essas duas variáveis aumentam, a tendência é observar um *IDHM* cada vez menor. Da mesma maneira, há uma conexão entre *IDHM* baixo e o aumento no número de *favelas*.

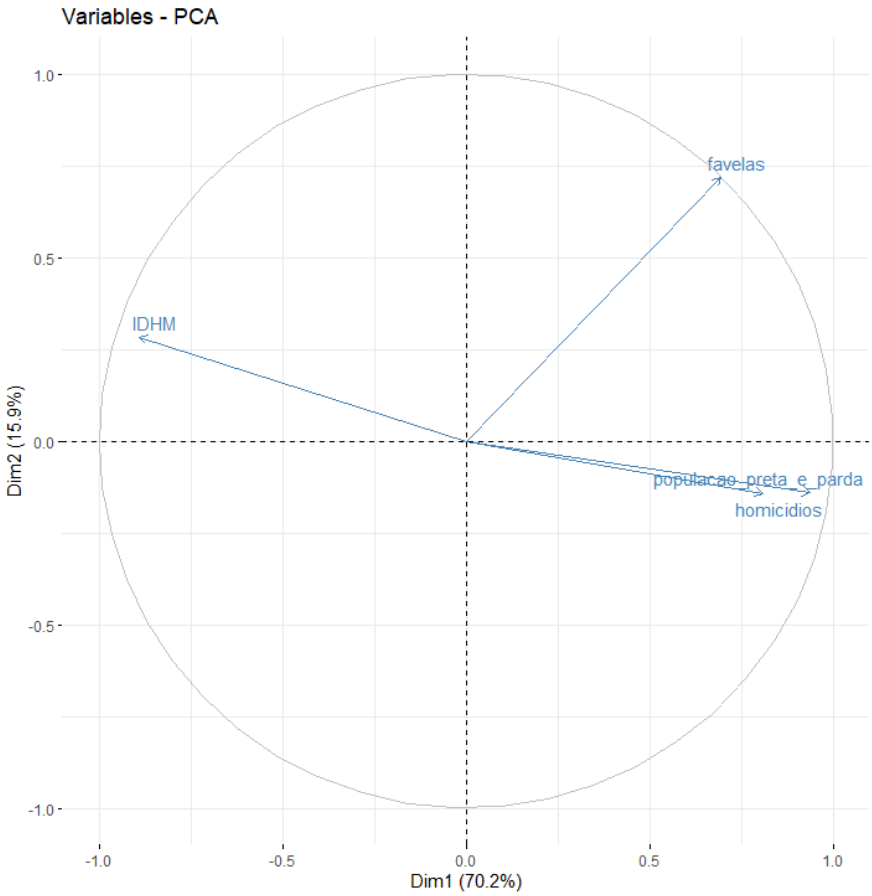

Figura 10: Gráfico PCA e variáveis

Antes de dar início a última etapa do trabalho, foi criado um gráfico com os bairros de São Paulo agrupados pelas regiões da cidade (Centro, Leste, Oeste, Norte e Sul). A linha de código $grupo <- as.factor(dados\_bairros\_completo\$Regiao$ combina os bairros por região. Para visualizar, a função $fviz\_pca\_biplot$ ilustra a distribuição das regiões no gráfico de PCA (Figura 11).

A zona sul apresenta alguns bairros com alto índice de *favelas* e ao mesmo tempo traz bairros com alto *IDHM.* Semelhantemente, há um grande contraste na variação de *homicidos* e *populacao_preta_e_parda* na zona leste. A partir do gráfico, fica evidente que boa parte dos bairros pertencentes a mesma região possuem características dissonantes. Logo, a aplicação de políticas sociais baseadas nas similaridades entre regiões da cidade de São Paulo não se apresentaria como o método mais eficiente e eficaz.

Alternativamente, a elaboração de clusters pode ser mais adequada para associar os bairros que, de fato, possuam correlações entre si.

Figura 11: Gráfico PCA por regiões de São Paulo

2.5 - **Análise de Cluster**

Análise de clusters é uma técnica de análise estatística usada para agrupar dados em grupos semelhantes, com base em atributos específicos. Em outras palavras a análise de cluster “tem por objetivo principal a alocação de observações em uma quantidade relativamente pequena de agrupamentos homogêneos internamente e heterogêneos entre si e que representem o comportamento conjunto das observações a partir de determinadas variáveis.” (Fávero e Belfore, 2017)

Conforme verificado na análise PCA, existe uma disparidade entre os parâmetros avaliados dentro das próprias regiões da cidade de São Paulo. Assim, para uma compreensão melhor dos bairros que apresentam relações similares entre suas variáveis será adotada a técnica de clusterização.

Primeiro, será criada uma matriz de dissimilaridade utilizando o método euclidiano. Em seguida, será aplicado o método *complete linkage* para a geração dos cluster. A figura 12 ilustra o dendrograma com a divisão dos clusters.

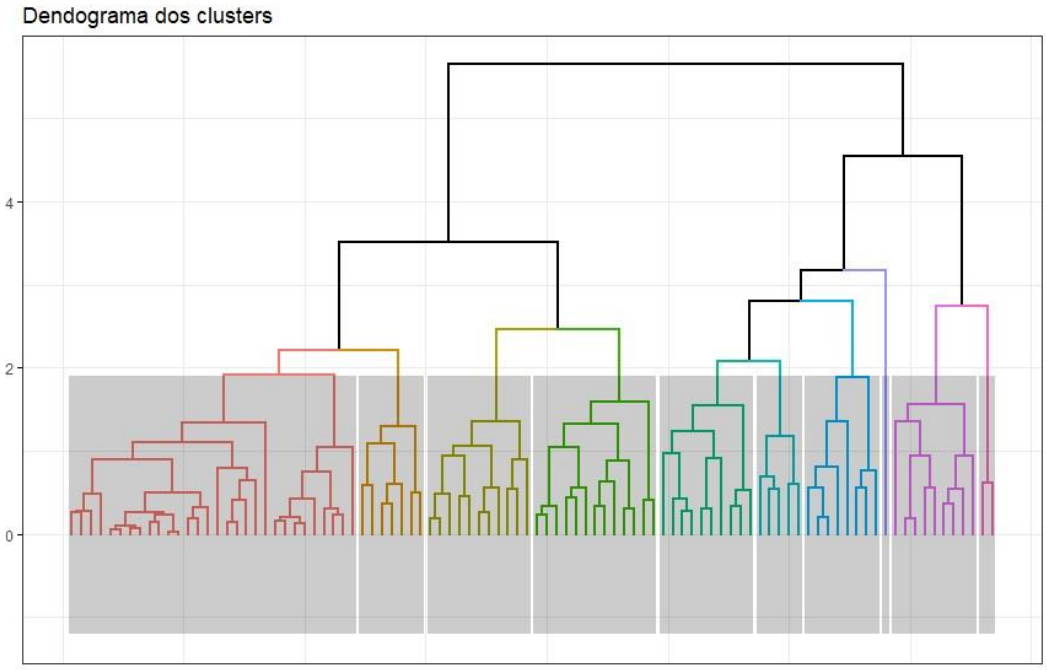

Figura 12: Dendograma dos clusters

Abaixo, os clusters definidos são plotados no mapa da cidade de São Paulo conforme a Figura 13.

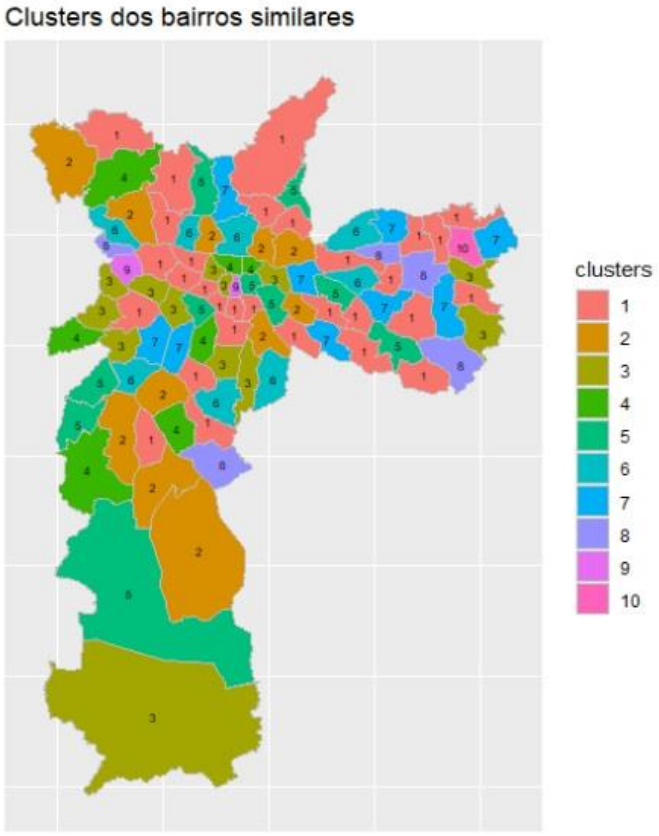

Figura 13: Clusters dos bairros similares

A partir da visualização dos bairros em clusters (Figura 14) é possível verificar que bairros próximos ao centro e de alto padrão, como Alto de Pinheiros, Perdizes, Bela Vista e Santa Cecília, tendem a ficarem no cluster 1. Bairros do centro histórico aparecem no cluster 3 com características similares a bairros periféricos do cluster 2.

A zona sul e a zona norte trazem a maioria de seus bairros tradicionais representados nos clusters 4 e 6. Em contraste, a proximidade entre bairros periféricos é evidenciada nos clusters 5, 7 e 8.

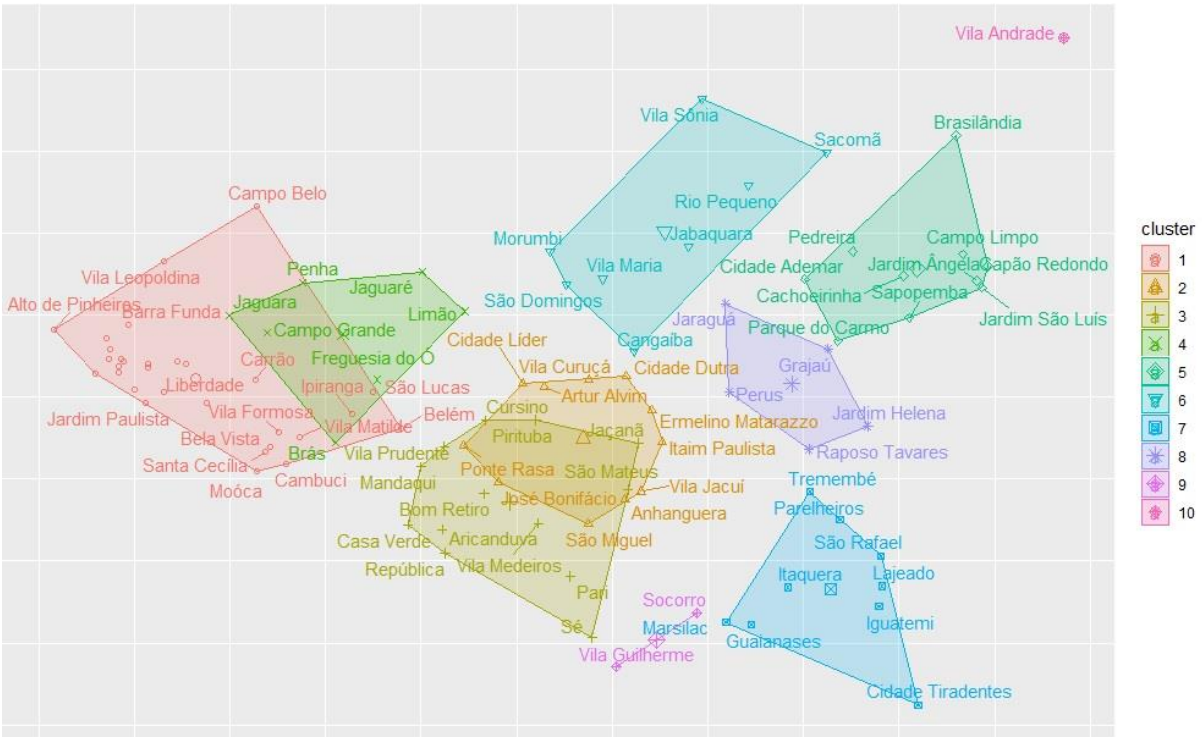

Figura 14: Clusters dos bairros

3 - **Considerações Finais**

O incentivo para a execução do presente trabalho deu-se a partir da constatação de que o principal instrumento de avaliação da situação de vida da população brasileira (Censo Demográfico – IBGE) não endereça concretamente alguns dados perenes de desigualdade, como a distribuição da população por cor da pele, localização de favelas e a violência nas cidades.

A discrepância na qualidade de vida baseada somente na cor da pele demonstra que o racismo estrutural vem se perpetuando desde o período colonial brasileiro.

Os dados apresentados indicam que a população preta e parda está mais exposta à violência e a condições menos favoráveis de moradia na cidade de São Paulo. No entanto, quando a separação dos bairros é feita por similaridades através da análise de clusters, não há homogeneidade.

Assim, devida à complexidade do processo histórico desordenado que conduziu o crescimento urbano da cidade de São Paulo, faz-se necessário um estudo minucioso em cada bairro para o estabelecimento de políticas públicas efetivas que visem a melhoria das condições de vida.

Referências

ALMEIDA, Silvio. **O que é racismo estrutural?** Belo Horizonte: Letramento, 2018.

ATLAS Brasil. **Atlas do Desenvolvimento humano no Brasil 2013**. Disponível em: <http://www.atlasbrasil.org.br>. Acesso em: 8 fev. 2023.

FÁVERO, Luiz Paulo Lopes e BELFIORE, Patrícia Prado. **Manual de análise de dados: estatística e modelagem multivariada com excel, SPSS e stata**. Rio de Janeiro: Elsevier, 2017.

FERNANDES, Florestan. **A sociologia no Brasil.** Petrópolis: Vozes, 1977.

IBGE. **Censo 2010.** Disponível em: <www.censo2010.IBGE.gov.br>. Acesso em: 8 fev. 2023.

MAPA DA DESIGUALDADE. Rede Nossa São Paulo, 2021. Disponível em: <https://www.nossasaopaulo.org.br/>. Acesso em: 8 fev. 2023.

O que é o IDHM. **UNDP**, 2023. Disponível em: <https://www.undp.org/pt/brazil/o-que-é-o-idhm>. Acesso em: 8 fev. 2023.

PINTO, Daniela Gomes Coordenação; COSTA, Marco Aurélio. Coordenação; MARQUES, Maria Luiza de Aguiar Coordenação. **O índice de desenvolvimento humano municipal brasileiro.** 2013.

SAMPAIO, Américo. **A geografia da desigualdade na cidade de São Paulo**. Revista Parlamento e Sociedade, São Paulo, v. 6, n. 10, p. 61-76, jan./jun. 2018. Disponível em: <https://parlamentoesociedade.emnuvens.com.br/revista/article/view/14>. Acesso em: 8 fev. 2023.