UNIVERSIDADE FEDERAL DE ALFENAS

LAILA BIANCA MEIRA MARTINS GARCIA

PNAD 2015: análise descritiva e de agrupamento dos domicílios da Região Sudeste do Brasil

VARGINHA/MG

2022

LAILA BIANCA MEIRA MARTINS GARCIA

PNAD 2015: análise descritiva e de agrupamento dos domicílios da Região Sudeste do Brasil

Trabalho de Conclusão do Programa Integrado de Pesquisa, Ensino e Extensão (PIEPEX) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência e Economia pela Universidade Federal de Alfenas.
Orientadora: Patrícia de Siqueira Ramos

VARGINHA/MG
2022

LAILA BIANCA MEIRA MARTINS GARCIA

PNAD 2015: análise descritiva e de agrupamento dos domicílios da Região Sudeste do Brasil

A banca examinadora abaixo-assinada, aprova o trabalho de conclusão do PIEPEX (TCP), apresentado como parte dos requisitos para a obtenção do grau de Bacharel Interdisciplinar em Ciência e Economia pela Universidade Federal de Alfenas.

Aprovada em: __ / __ / ____

Prof.ª Dr.ª Patrícia de Siqueira Ramos
Universidade Federal de Alfenas

Prof.ª Dr.ª Gislene Araújo Pereira
Universidade Federal de Alfenas

Prof.ª Dr.ª Luciene Resende Gonçalves
Universidade Federal de Alfenas

RESUMO

Este trabalho tem como objetivo realizar uma análise descritiva dos dados de domicílios da Região Sudeste do Brasil, divulgados pela Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2015, e um agrupamento de domicílios com perfis semelhantes pertencentes a essa região. Com 85 variáveis à disposição, foi feita uma limpeza inicial dos dados e uma análise descritiva, por meio de tabelas, gráficos e medidas resumo. Em seguida, foram selecionadas quatro variáveis quantitativas para realizar o agrupamento, sendo elas: total de moradores, número de cômodos dos domicílios, número de cômodos servindo de dormitório e rendimento mensal domiciliar *per capita*. Para a análise de agrupamento foram utilizados o método multivariado hierárquico de Ward, para definir o número de grupos, e o não hierárquico das *k*-médias, para o agrupamento em si. A linguagem de programação *python* foi utilizada em todas as etapas do trabalho. O método de Ward sugeriu a divisão dos domicílios em quatro grupos e, após a aplicação das *k*-médias, obteve-se um resumo estatístico das variáveis por grupo, traçando-se assim quatro perfis de domicílios da Região Sudeste com base nas variáveis analisadas. Houve destaque, principalmente, de dois grupos de domicílios com realidades opostas. Um deles apresentou alto valor médio de rendimento mensal *per capita* (R$9.022,50), poucos moradores e mais cômodos no domicílio. O outro apresentou rendimento mensal domiciliar *per capita* médio de R$8.038,22 a menos, com uma média maior de moradores e menos cômodos no domicílio.

Palavras-chave: domicílios; análise multivariada; agrupamento, Ward; *k*-médias.

SUMÁRIO

1- INTRODUÇÃO

Em países com dimensões continentais, como o Brasil, e com realidades socioeconômicas muito distintas entre a própria população, é necessário um levantamento de dados e variáveis que geram diversos impactos nas condições familiares. Para a geração dessas informações foi criada, pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 1967, a Pesquisa Nacional por Amostra de Domicílios (PNAD). A PNAD tem a finalidade de realizar uma pesquisa básica que se estende nacionalmente, com uma amostra probabilística de domicílios particulares e coletivos ocupados, na qual são gerados continuamente dados sobre trabalho, rendimento, população, habitação e educação (IBGE, 2016).

Com o intuito de que haja um estudo simultâneo de inúmeras variáveis, como as disponibilizadas pela PNAD, a estatística multivariada é extremamente importante e facilitadora para que sejam realizadas essas análises em um ambiente complexo, com a finalidade de clarear e reduzir as estruturas e dados em análise, classificar e agrupar, buscar padrões e ser um meio facilitador na elaboração de hipóteses e, por fim, testá-las (JOHNSON; WICHERN, 1992 apud BAKKE; LEITE; SILVA, 2008). Uma das técnicas multivariadas mais populares é a análise de agrupamento, em que o "objetivo é formar grupos com propriedades homogêneas de amostras heterogêneas grandes" (HÄRDLE; SIMAR, 2007 apud BAKKE; LEITE; SILVA, 2008).

Deste modo, o objetivo deste artigo é realizar uma análise descritiva dos dados dos domicílios da região Sudeste do Brasil, divulgados pela PNAD 2015, com o intuito de observar e descrever tendências e padrões. Além disso, será realizada uma análise de agrupamento para identificar grupos de domicílios com perfis semelhantes no que se refere à situação desses domicílios em relação à renda e moradia.

O trabalho está estruturado em cinco seções. Após esta introdução é apresentado o referencial teórico, com uma apresentação dos principais pontos da PNAD e, em seguida, da análise de agrupamento. Na terceira seção é exposta a metodologia utilizada, com ênfase em três pontos, sendo o primeiro uma explicação dos dados empregados, o segundo uma análise descritiva e o terceiro uma análise de agrupamento. A quarta seção contempla os resultados e discussões sobre as análises realizadas e, por fim, na quinta seção, as considerações finais.

2- REFERENCIAL TEÓRICO

2.1 - PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS (PNAD - IBGE)

No início da década de 1960, surgiu a necessidade da criação de um planejamento para o desenvolvimento socioeconômico dos países latinos e, a partir de um modelo de pesquisa idealizado pela *United States Agency for International Development* (USAID) e pela *Inter-American Statistical Institute* (IASI), inicia-se um processo de cooperação internacional para a pesquisa contínua dessas estatísticas. No Brasil, o processo de criação e execução de um modelo de pesquisa foi elaborado pelo IBGE em 1966 e sua primeira edição ocorreu em 1967, sendo denominada Pesquisa Nacional por Amostra de Domicílios (IBGE, 2015).

A PNAD surge com o objetivo de realizar uma pesquisa básica que abrange todo o país, com sua amostragem domiciliar e de periodicidade variável, discorre dados que se projetam como possíveis meios de criação de políticas públicas que podem resolver de forma realísticas os problemas socioeconômicos do Brasil. As principais características e variáveis da PNAD são apresentadas na Tabela 1.

Tabela 1 - Características e variáveis utilizadas na PNAD

Características	Variáveis
Demográficas e sociais	Sexo, cor, condição na unidade domiciliar, posição na família e no domicílio, número na família e data de nascimento dos moradores.
Educacionais	Alfabetização, escolaridade (série e grau frequentados) e nível de instrução das pessoas que não são estudantes (última série concluída e grau correspondente).
Mão de obra	- Para as pessoas de 10 anos de idade ou mais: condição de atividade. - Para as pessoas ocupadas: ocupação, atividade e posição na ocupação no trabalho principal, horas normalmente trabalhadas por semana no trabalho principal e nos outros trabalhos, e se é contribuinte de instituto de previdência pelo trabalho. - Para as pessoas desocupadas: tempo de procura de trabalho, ocupação, atividade, posição na ocupação e motivo da saída, se recebeu fundo de garantia, e tempo de permanência em relação ao

	último trabalho remunerado.
Rendimentos	Rendimento mensal normalmente recebido do trabalho principal e dos outros trabalhos, aposentadoria, pensão, abono de permanência, aluguel e outros rendimentos.
Habitação	- Espécie de domicílio - Para os domicílios particulares permanentes: tipo, estrutura, abastecimento de água, esgotamento sanitário, uso de instalação sanitária, destino do lixo, iluminação elétrica, número de cômodos, condição de ocupação, aluguel ou prestação mensal, filtro de água, fogão, geladeira, rádio e televisão.

Fonte: IBGE (1991)

O foco deste trabalho são os municípios da Região Sudeste do Brasil, então esta será a área a ser tratada. Ao longo das últimas décadas, as condições socioeconômicas dessa região do país vêm sofrendo mudanças constantemente e um dos principais pontos para entender essas transformações é analisar profundamente os domicílios do local. Segundo o Levantamento de Informações Territoriais (LIT, 2010 apud UOL, 2013), no ano de 2010, a Região Sudeste concentrava 49,8% dos domicílios localizados em favelas, sendo que 31,6% tinham um rendimento domiciliar per capita de até meio salário mínimo e 0,9% tinham mais de cinco salários mínimos. A PNAD é uma das pesquisas mais importantes para que frequentemente essas estatísticas sejam atualizadas e assim haja clareza das mudanças que vêm ocorrendo em diversas variáveis, como as contempladas na PNAD.

2.2 - ANÁLISE MULTIVARIADA

Para Hair et al. (2009), a análise multivariada pode ser definida como um conjunto de métodos estatísticos que analisa de forma síncrona inúmeras variáveis em busca de contribuir para a melhor tomada de decisão possível. No cotidiano é cada vez mais necessário o uso deste processo, para obter uma visão mais ampla dos problemas e assim correr menos riscos, podendo ser utilizada por empresas, órgãos governamentais e centros de pesquisas.

Os dados multivariados podem ser apresentados em forma matricial, em que uma amostra aleatória com tamanho *n* (linhas ou observações) possui *p* (colunas) variáveis, criando-se uma matriz **X** de dimensão *n* x *p*, como apresentado a seguir:

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \dots & X_{pp} \end{bmatrix},$$

Esses tipos de dados apresentam grande complexidade, e algumas questões podem ser levantadas ao se observar uma amostra de dados multivariados. Uma delas envolve a tentativa de se dividir a amostra em grupos com observações similares entre si ou com variáveis com características parecidas. Há algumas questões relevantes envolvendo esse tema:

1) Como as *p* variáveis se relacionam dentro de cada grupo?

2) Os grupos diferem significativamente quanto aos valores médios das variáveis?

3) Os grupos mostram quantidades similares de variação para as variáveis?

4) Caso os grupos sejam diferentes em termos de distribuição das variáveis, é possível construir alguma função destas variáveis que separe dois grupos?

Para responder às questões acima foram desenvolvidas diversas técnicas multivariadas, cujos objetivos são diferentes, como por exemplo a análise fatorial, análise de componentes principais, regressão múltipla, análise conjunta, correlação canônica, escalonamento e análise de agrupamento (MANLY, 2005 apud SONG, 2013).

A análise de agrupamento tem como objetivo reunir o máximo de informações possíveis dos objetos (indivíduos, elementos) buscando semelhanças entre as variáveis que as tornem um grupo e buscando diferenças entre estes grupos (VICINI, 2005). Este processo é realizado em três etapas, sendo o primeiro a busca por uma medida que mostre as similaridades entre as variáveis e quantos grupos

serão formados, o segundo passo é o processo de agrupamento e por último deve-se formar um perfil das variáveis para estabelecer sua composição (HAIR et al, 2009).

Existem dois modos de realizar a análise de agrupamento, são eles o método hierárquico e o método não hierárquico. Segundo Seidel (2008), o método hierárquico busca criar uma associação das variáveis a partir da formação de uma estrutura em forma de árvore (dendrogramas). Existem alguns tipos de métodos hierárquicos, os principais são: vizinho mais próximo, vizinho mais distante, distância média, centróide e Ward. Além disso, alguma medida de distância deve ser adotada, de forma a representar a similaridade entre as observações, sendo a mais comum a distância euclidiana.

De acordo com Hair et al (2009), o método de agrupamento hierárquico Ward consiste em calcular a soma dos quadrados de todas as variáveis dos agrupamentos e usar isso como um medidor de similaridade para uni-los, tendo-se, dessa forma, grupos com observações semelhantes entre si. Esse método tende a formar agrupamentos balanceados no que se refere à quantidade de observações, ou seja, o número de elementos dentro de cada grupo tende a ser similar. Por esse motivo, esse é um dos métodos hierárquicos mais utilizados

O agrupamento não hierárquico, diferentemente do hierárquico, necessita que haja a definição da quantidade de grupos anteriormente ao processo de agrupamento, buscando ter o maior número possível de pontos semelhantes entre as variáveis de um mesmo grupo, prezando pela alta precisão (HAIR et al, 2009). O método principal e mais usado deste modelo é o das *k*-médias, que consiste na divisão em *k* grupos, e o centróide de cada agrupamento é calculado a partir da soma das distâncias ao quadrado do centróide de cada objeto do grupo. O que se busca é a minimização dessa soma de quadrados (LATTIN, 2003).

Na próxima seção serão apresentados os dados e os métodos adotados para analisá-los, incluindo os passos da análise de agrupamento.

3 - METODOLOGIA

Foi utilizada a edição do ano de 2015 da PNAD, realizada nas datas de referência de 27 de setembro de 2014 e 26 de setembro de 2015, em domicílios

particulares e coletivos ocupados na zona urbana e rural, com uma amostra de 356.904 pessoas e 151.189 domicílios em toda a extensão federativa.

O foco desta pesquisa foram os dados referentes a domicílios da zona urbana da Região Sudeste, retirados da PNAD 2015 (IBGE, 2022). Esses dados foram analisados por meio da linguagem de programação *Python* (PYTHON, 2017) e da interface *Google Colab*. Foram realizadas uma análise descritiva e uma busca de padrões e tendências entre elas, por meio da análise de agrupamento.

Das 85 variáveis de domicílio apresentadas pela PNAD 2015, este trabalho teve enfoque em quatro para o agrupamento. Essa seleção foi feita porque a maioria das variáveis eram de natureza qualitativa, não sendo possível utilizá-las para a análise de agrupamento, que exige variáveis quantitativas. Sendo assim, as variáveis selecionadas foram:

1. Total de moradores;
2. Número de cômodos dos domicílios;
3. Número de cômodos servindo de dormitório;
4. Rendimento mensal domiciliar *per capita*.

Dessa forma, de acordo com a classificação apresentada na Tabela 1, as variáveis selecionadas são de habitação e rendimento. Inicialmente, os dados apresentavam 151.189 domicílios (de toda a extensão federativa), após a seleção dos domicílios da zona urbana da Região Sudeste restaram 45.546 domicílios. Por último, os domicílios sem declaração (para a maioria das variáveis é possível para o respondente escolher a opção “sem declaração”) foram retirados, chegando-se a 36.326 domicílios. Em relação às variáveis, inicialmente eram 85, logo em seguida foram retiradas as usadas para controle da amostragem, ficaram 61 variáveis. Na seleção final foram escolhidas variáveis quantitativas, chegando a quatro variáveis usadas para o agrupamento. Ao final, tem-se um total de 36.326 domicílios e quatro variáveis para serem analisadas, ou seja, a dimensão resultante dos dados (n x p) foi, então, n = 36.326 observações e p = 4 variáveis. As bibliotecas do *Python* utilizadas foram: *pandas* (para ler e tratar o conjuntos de dados); *numpy* e *scipy* (para os cálculos dos vetores e matrizes); *matplotlib* e *seaborn* (para a criação de gráficos); *scipy.cluster.hierarchy, scipy.spatial.distance* e *sklearn.cluster* (para a análise de agrupamento).

A análise descritiva foi feita por meio do cálculo de medidas resumo, gráficos e tabelas. No resumo estatístico das variáveis utilizaram-se valores como média, mediana, valor máximo, valor mínimo, quantil 25% (q_25), quantil 75% (q_75) e desvio padrão. Além dessas medidas, também foram obtidas contagens simples, e gráficos de colunas e histogramas para uma melhor visualização.

Na análise de agrupamento, foram aplicados métodos hierárquicos e o não hierárquico das *k*-médias. Dentre os métodos hierárquicos foram utilizados cinco, sendo eles: vizinho mais próximo, vizinho mais distante, centróide, distância média e Ward. Houve uma melhor disposição dos domicílios com a técnica de Ward, com grupos mais balanceados em relação ao número de observações por grupo. Por essa razão, e também por ele ser um dos métodos hierárquicos mais utilizados na prática, ele foi o escolhido. A escolha da quantidade de grupos de domicílios foi elaborada através de um corte no dendrograma, com a finalidade de encontrar o maior afastamento possível entre os grupos, assim foi definido que a divisão ocorreria em quatro grupos.

Após a definição do número de grupos, o agrupamento em si foi realizado por meio da técnica das *k*-médias com *k*=4. Depois da aplicação do método, os valores da média e mediana para cada grupo de domicílios foram calculados. Dessa forma, foi possível traçar o perfil desses quatro agrupamentos de domicílios do Sudeste do Brasil.

4- ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 - ANÁLISE DESCRITIVA

Nesta seção será apresentada a análise descritiva das quatro variáveis selecionadas para o agrupamento, sendo elas: total de moradores, número de cômodos dos domicílios, número de cômodos servindo de dormitório e rendimento mensal domiciliar *per capita*, e também a variável tipo de domicílio.

Uma informação interessante é que os tipos de domicílios da Região Sudeste mais presentes são as casas, com 83,2% dos domicílios, logo em seguida estão os apartamentos, em uma porcentagem de 16,6%, e com menor expressão aparecem os cômodos com 0,2%. A Tabela 2 apresenta o resumo estatístico das variáveis definidas para a análise dos domicílios da região Sudeste.

Tabela 2 - Resumo estatístico das variáveis socioeconômicas dos domicílios da região Sudeste

	Total de Moradores	N° de cômodos dos domicílios	N° de cômodos servindo de dormitório	Rendimento mensal domiciliar *per capita*
Média	2,90	5,83	1,80	1.485,35
Desvio padrão	1,42	2,05	0,77	2.361,23
Valor mínimo	1,00	1,00	1,00	0,00
q_25	2,00	5,00	1,00	540,00
mediana	3,00	5,00	2,00	887,00
q_75	4,00	7,00	2,00	1.550,00
Valor máximo	16,00	27,00	7,00	150.000,00

Fonte: Elaboração própria, a partir dos dados da PNAD (2015).

A partir da análise da Tabela 2, percebe-se que: todos domicílios possuem ao menos um morador, sendo a média de 2,9 moradores por domicílio, enquanto a mediana tem um leve aumento, em relação a média, e mostrou que metade dos domicílios apresentaram 3 moradores ou menos, a quantidade máxima foi de 16 moradores por domicílio. Ao fazer uma contagem simples foi visto que a maior parte dos domicílios possuem dois moradores, com 26,7%, em seguida, 26,1% dos domicílios têm três moradores, seguidos por quatro moradores, com 19,3%.

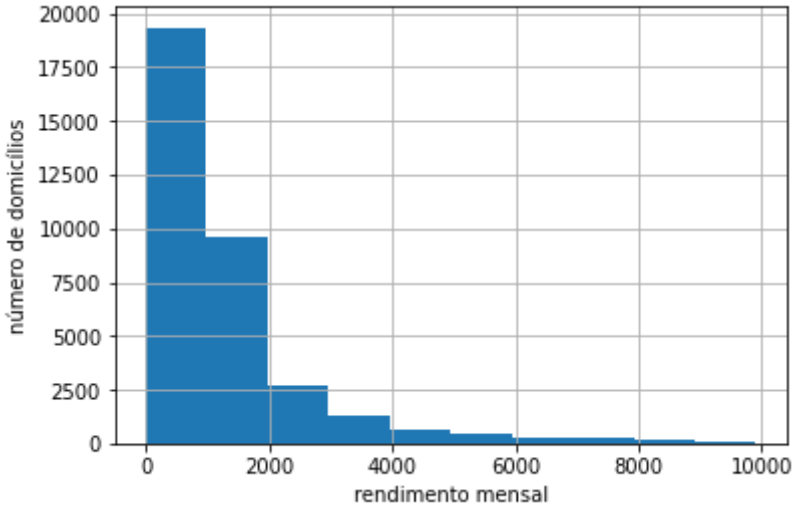
Em relação ao número de cômodos dos domicílios, foi observado que os domicílios têm pelo menos um cômodo, sendo a média de 5,83 cômodos por domicílio, enquanto a mediana é menor que a média com 5 cômodos por domicílio, o valor máximo chega à quantidade de 27 cômodos por domicílio. A quantidade de cômodos que aparecem mais vezes na pesquisa são 5 cômodos com 29,8%, seguidos por 6 cômodos com 19,4% e, em terceiro, com 14,5%, foram os domicílios com 4 cômodos.

Ao observar a variável número de cômodos servindo de dormitório, tem-se ao menos um cômodo servindo de dormitório, sendo a média de 1,80 cômodos por

domicílio, enquanto pelo menos metade dos domicílios do Sudeste possui dois cômodos ou menos servindo de dormitório, sendo o valor máximo de cômodos utilizados como dormitório, 7. Na contagem da variável, dois cômodos servindo de dormitório aparecem em primeiro lugar, com 43,5%, seguido por um cômodo com 39,4% e, em terceiro, aparecem três cômodos, com 15,11%.

Ao analisar a variável rendimento mensal domiciliar *per capita*, calculado a partir da divisão dos rendimentos mensais do domicílio pelo número de moradores, observa-se que o valor mínimo encontrado foi de R$0,00. A média dos domicílios foi de R$1.485,35, enquanto a mediana é bem menor do que a média, apresentando um valor de R$887,00. O valor máximo encontrado foi de R$150.000,00. Como forma de ilustrar a distribuição dos valores de rendimento mensal *per capita* foi construído um histograma. Porém, como essa variável apresenta alguns valores discrepantes, muito maiores do que os demais, para que fosse possível observar tal distribuição, foi feita uma seleção dos dados de forma a ilustrar apenas valores abaixo de R$10.000,00 mensais. Tal distribuição de frequências é apresentada no histograma do Gráfico 1.

Gráfico 1 - Histograma com a distribuição de frequência do rendimento mensal domiciliar *per capita* dos domicílios da Região Sudeste do Brasil, considerando-se apenas valores abaixo de R$10.000,00 mensais

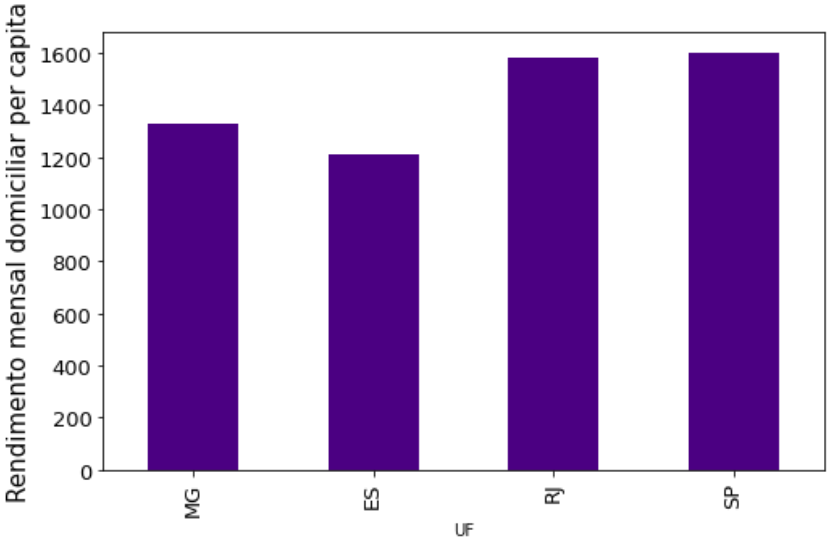


Fonte: Elaboração própria, a partir dos dados da PNAD (2015).

Percebe-se, a partir do histograma do Gráfico 1, que há uma concentração muito grande de domicílios com renda domiciliar *per capita* mensal menor do que R$2.000,00. E ainda é possível notar que a maior frequência se encontra em valores abaixo de R$1.000,00. Isso já havia sido mostrado com os dados presentes na Tabela 2.

Foi obtido um gráfico de colunas com os valores médios de rendimento mensal *per capita* por estado da Região Sudeste (Gráfico 2).

Gráfico 2 - Média do rendimento mensal domiciliar *per capita* dos estados da Região Sudeste



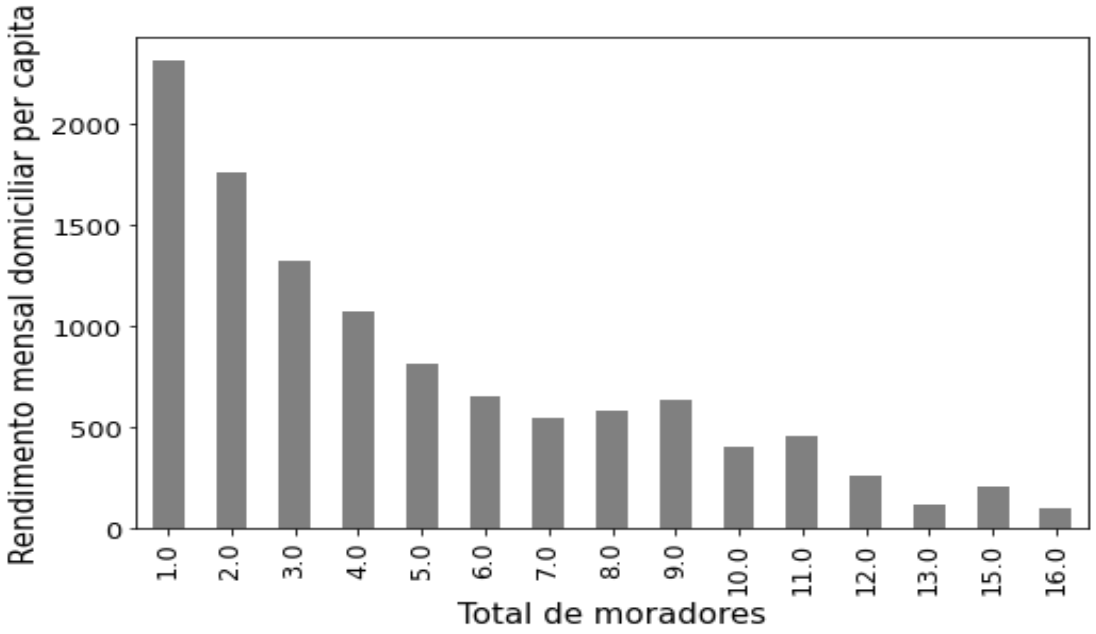
Fonte: Elaboração própria, a partir dos dados da PNAD (2015).

De acordo com o Gráfico 2, é possível perceber que o estado de São Paulo possui a maior média de rendimento mensal domiciliar *per capita* da região Sudeste, com o valor de R$1.601,79, seguido do Rio de Janeiro com R$1.579,04, depois vem Minas Gerais com R$1.326,44 e por último Espírito Santo com R$1.210,06. Como há uma diferença do estado com maior média (SP) de rendimento mensal domiciliar *per capita* para o menor (ES) de R$391,73, mostrando que existe um certo distanciamento entres os estados do Sudeste em relação a esta variável.

Ao relacionar as variáveis total de moradores e rendimento mensal domiciliar *per capita*, tem-se que domicílios com um morador possuem o maior rendimento mensal médio, sendo este de R$2.313,26; em segundo lugar estão os domicílios

com dois moradores que possuem um rendimento de R$1.760,68; logo em seguida, três moradores por domicílio apresentam média de R$1.318,77. Tais valores médios são apresentados no Gráfico 3.

Gráfico 3 - Rendimento mensal domiciliar *per capita* por total de moradores dos domicílios



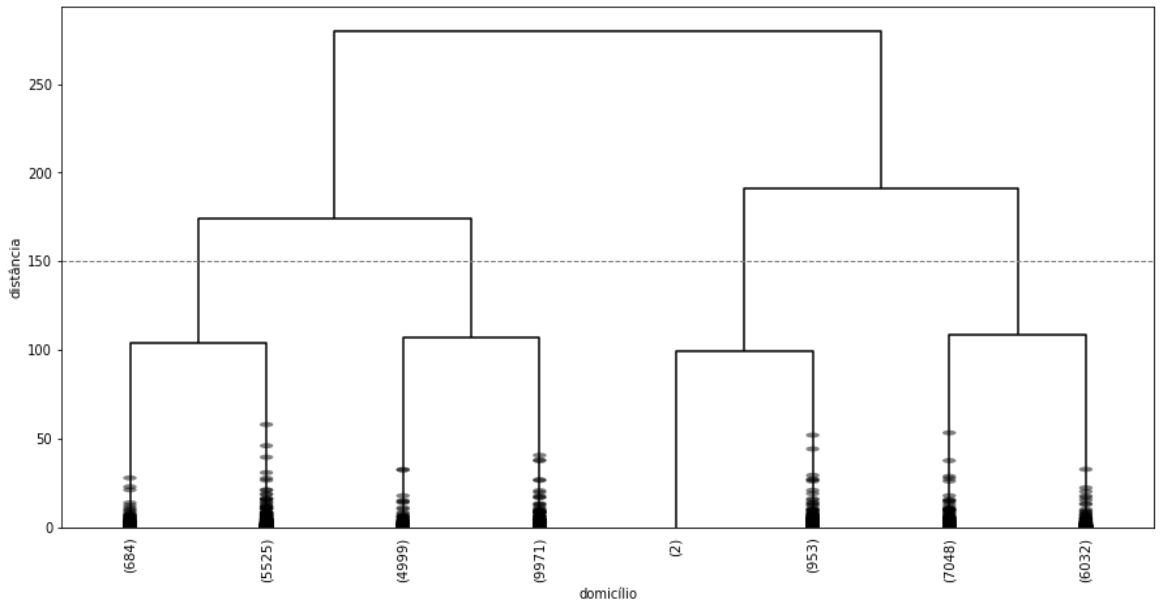
Fonte: Elaboração própria, a partir dos dados da PNAD (2015).

4.2- ANÁLISE DE AGRUPAMENTO

4.2.1 - Método Ward

Para a análise de agrupamento hierárquico e definição do número de grupos *k* foi utilizado o método Ward. A Figura 2 mostra o resultado do dendrograma deste método, apresentando um corte em quatro grupos de domicílios da Região Sudeste. Esse corte foi escolhido pois possui a maior área livre de separação entre os grupos.

Figura 2 - Dendrograma dos grupos de domicílios pelo método Ward



Fonte: Elaboração Própria, a partir dos dados da PNAD (2015)

No dendrograma de grupos de domicílios, como mostrado na Figura 2, foi realizado um corte em quatro grupos, o primeiro grupo contém 6.209 domicílios, o segundo 14.970, o terceiro 955 e o quarto 13.080. Este método foi utilizado para uma melhor observação de como ocorreu o afastamento entre os grupos, e assim pode-se definir a melhor quantidade de grupos para a análise de agrupamento, neste caso *k=4*.

4.2.2 - Método da *k-médias*

Como definido anteriormente pelo método Ward, o número de grupos utilizados para o método das *k-médias* foram 4. A distribuição ficou da seguinte forma: grupo 0 contendo 12.188 domicílios, grupo 1 com 14.857, grupo 2 com 1.475 e o grupo 3 com 6.694. Na Tabela 3 são apresentadas as médias e medianas dos grupos citados acima.

Tabela 3 - Médias e medianas dos grupos pelo método *k-médias*.

		Total de moradores	N° de cômodos dos domicílios	N° de cômodos servindo de dormitório	Rendimento mensal domiciliar *per capita*
Grupo 0	Média	1,69	4,93	1,00	1.351,28
	Mediana	2,00	5,00	1,00	1.000,00
Grupo 1	Média	3,29	5,37	1,95	984,28
	Mediana	3,00	5,00	2,00	775,00
Grupo 2	Média	2,01	9,25	1,50	9.022,50
	Mediana	2,00	9,00	1,00	7.537,00
Grupo 3	Média	4,42	7,63	2,95	1.180,95
	Mediana	4,00	7,00	3,00	850,00

Fonte: Elaboração própria, a partir dos dados da PNAD (2015)

De acordo com a análise da Tabela 3, é possível perceber que o grupo 0 se destaca com a menor média de total de moradores, sendo de 1,69, número de cômodos dos domicílios, com 4,93, e de número de cômodos servindo de dormitório, com 1,00, sendo que a mediana das 3 variáveis não sofrem grandes mudanças em relação à média. A média de rendimento mensal domiciliar *per capita* é a segunda maior com R$1.351,28, ficando atrás apenas do grupo 2, a mediana teve uma queda de R$351,28, deste modo metade dos domicílios deste grupo recebem R$1.000,00 ou menos.

Os domicílios do grupo 1 possuem a pior média de rendimento mensal domiciliar *per capita*, sendo de R$984,28, a mediana é inferior à média tendo o valor de R$775,00. Pertence a este grupo a segunda maior média de total de moradores, ficando atrás apenas do grupo 3, com 3,29 moradores por domicílio, a mediana teve uma pequena diminuição em relação a média; possui também a segunda maior média de número de cômodos servindo de dormitório, sendo inferior ao grupo 3,

com 1,95 cômodos. Em relação ao número de cômodos dos domicílios, apresenta a segunda pior média com 5,37 cômodos por domicílio.

Ao grupo 2 pertence a maior média de rendimento mensal domiciliar *per capita* dos grupos da região Sudeste, com um valor extremamente alto de R$ 9.022,50, mais do que 7 vezes o valor da segunda maior média (grupo 0), a mediana sofre uma queda brusca, em relação a média, de R$1.485,50 e passa a ser de R$7.537,00. Este grupo possui também a maior média de número de cômodos por domicílios, com 9,25 cômodos. Em relação a média de total de moradores, tem o segundo pior índice entre os grupos,ficando atrás apenas do grupo 0, com 2,01 por domicílio; também tem a segunda pior média da variável número de cômodos servindo de dormitório, novamente atrás do grupo 0, com 1,50.

O grupo 3 se destaca pela maior média de total de moradores, com 4,42 por domicílio; também possui a maior média de número de cômodos servindo de dormitório, com 2,95 cômodos. Possui a segunda maior média de número de cômodos dos domicílios, atrás apenas do grupo 2, com 7,63 cômodos por domicílio. Em relação ao rendimento mensal domiciliar *per capita* tem a segunda pior média, sendo superior apenas ao grupo 1, com um valor de R$1.180,95, a mediana tem uma queda, em relação a média, de R$330,95, tendo assim em metade dos seus domicílios rendimentos iguais ou menores a R$850,00.

5 - CONSIDERAÇÕES FINAIS

Este trabalho teve como finalidade apresentar uma análise descritiva dos domicílios da região Sudeste do Brasil, para isso foram usados os dados da PNAD 2015. Desta forma, foi utilizado para a análise os métodos multivariados, assim permitindo que ocorresse uma análise de forma síncrona de diversas variáveis, gerando uma visão mais ampla da situação real dos domicílios. A análise de agrupamento foi a técnica multivariada escolhida para estes dados, tendo enfoque nos métodos de Ward e *k-médias*.

Ao analisar os domicílios da região Sudeste, é possível observar que metade deles possuem um rendimento mensal *per capita* igual ou menor a R$887,00, um valor R$99,00 maior que o salário mínimo, que no ano de 2015 era de R$788,00

(BRASIL, 2014). Em média, esses domicílios possuem 2,90 moradores, com 5,83 cômodos e tendo ao menos um desses cômodos servindo de dormitório.

Na análise de agrupamento foi feito um corte de quatro grupos de domicílios na região Sudeste. O grupo 2 se destaca com as melhores condições socioeconômicas, com uma média de rendimento mensal domiciliar *per capita* extremamente alta de R$9.022,50 com apenas 2,01 moradores e com uma alta quantidade de cômodos de 9,25. Já o grupo 1 vive uma realidade oposta, possuindo uma média de rendimento mensal domiciliar *per capita* com uma diminuição de R$8.038,22 em relação ao primeiro grupo citado, mas com uma média maior de moradores de 3,29 e com menos cômodos nos domicílios.

A realidade socioeconômica dos domicílios da Região Sudeste do Brasil é de extrema importância para todos os setores do país. Este estudo pode auxiliar, através da análise multivariada, trazendo informações para que o governo e as empresas possam compreender de forma ampla a realidade que cerca os domicílios, para que corram o menor risco possível nas decisões de seus projetos. Sugere-se que em futuras pesquisas seja feita a análise de outras variáveis, trazendo ainda mais informações a respeito dos domicílios dessa região.

REFERÊNCIAS BIBLIOGRÁFICAS

BAKKE, H.; LEITE, A.; SILVA, L. Estatística multivariada: aplicação da análise fatorial na engenharia de produção. **Revista Gestão Industrial,** Ponta Grossa, v.04, n.04, p.01-14, 2008. Disponível em: ESTATÍSTICA MULTIVARIADA: APLICAÇÃO DA ANÁLISE FATORIAL NA ENGENHARIA DE PRODUÇÃO | Bakke | Revista Gestão Industrial (utfpr.edu.br). Acesso em: 07 de Março de 2022.

BRASIL. **Decreto n. 8.381**, de 29 de dezembro de 2014. Dispõe sobre o valor do salário mínimo e a sua política de valorização de longo prazo. Disponível em: Decreto nº 8381 (planalto.gov.br). Acesso em: 31 de março de 2022.

HAIR, J.F. et al. **Análise multivariada de dados**. 6 ed. Porto Alegre: Bookman, 2009.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Para compreender a PNAD**: um texto simplificado. Rio de janeiro: IBGE, 1991.

_____.. **Pesquisa nacional por amostra de domicílios**: síntese de indicadores 2015. Rio de Janeiro: IBGE, 2016.

_____. **PNAD**: um registro histórico da pesquisa nacional por amostra de domicílios 1967-2015. Rio de Janeiro: IBGE, 2015.

_____. **Pesquisa Nacional por Amostra de Domicílios (PNAD) 2015**. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao . Acesso em: 20 jan. 2022.

LATTIN, J.; CARROLL, J.D.; GREEN, P.E. **Análise de dados multivariados**. São Paulo: Cengage Learning Edições LTDA, 2003).

PYTHON. **The Python programming language**. Disponível em: https://docs.python.org/3/reference. Acesso em: 28 de janeiro de 2022.

SEIDEL, E. et al. Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite. **Ciência e Natureza**, v.30, n.01, p.07-15. Disponível em: Comparação Entre o Método Ward e o Método k-médias no Agrupamento de Produtores de Leite | Ciência e Natura (ufsm.br). Acesso em: 17 de março de 2022.

SONG, F. **Técnicas de análise multivariada com aplicações a dados de natureza biológica**. TCC (Graduação em Ciências Biológicas), Universidade Estadual Paulista, Rio Claro, 2013.

UOL. Região Sudeste concentra metade dos domicílios em favelas do Brasil. **Uol**, 2013.Disponível em: https://noticias.uol.com.br/cotidiano/ultimas-noticias/2013/11/06/regiao-sudeste-concentra-metade-dos-domicilios-em-favelas-do-brasil.htm. Acesso em: 21 de Março de 2022.

VICINI, L. **Análise multivariada da teoria à prática**. Monografia (Pós-Graduação em Estatística e Modelagem Quantitativa), Universidade Federal de Santa Maria, Santa Maria, 2005.