UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

JONAS NASCIMENTO SANTOS

MODELAGEM DOS PREÇOS DE IMÓVEIS RESIDENCIAIS NO MUNICÍPIO DE ARACAJU VIA MISTURA DE MODELOS DE REGRESSÃO

São Cristóvão - SE

2025

JONAS NASCIMENTO SANTOS

MODELAGEM DOS PREÇOS DE IMÓVEIS RESIDENCIAIS NO MUNICÍPIO DE ARACAJU VIA MISTURA DE MODELOS DE REGRESSÃO

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística .

Orientador: Prof. Dr. Luiz Henrique Gama Dore de Araújo

São Cristóvão - SE

2025

JONAS NASCIMENTO SANTOS

MODELAGEM DOS PREÇOS DE IMÓVEIS RESIDENCIAIS NO MUNICÍPIO DE ARACAJU VIA MISTURA DE MODELOS DE REGRESSÃO

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística .

Aprovado em 01 de Abril de 2025.

Documento assinado digitalmente
gov.br
LUIZ HENRIQUE GAMA DORE DE ARAUJO
Data: 10/04/2025 15:16:50-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Luiz Henrique Gama Dore de Araújo
Orientador

Documento assinado digitalmente
gov.br
CLEBER MARTINS XAVIER
Data: 09/04/2025 17:38:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cleber Martins Xavier
Convidado 1

Documento assinado digitalmente
gov.br
CARLOS ROBSON SILVESTRE DA SILVA MARTIN
Data: 09/04/2025 15:13:43-0300
Verifique em https://validar.iti.gov.br

Carlos Robson Silvestre da Silva Martins
Convidado 2

São Cristóvão - SE
2025

Este trabalho é uma homenagem a todos os professores que me influenciaram no meu crescimento acadêmico

Agradecimentos

Gostaria de expressar minha profunda gratidão a Deus, por me fazer persistir todos esses anos, e principalmente à minha família, sobretudo aos meus pais, Gilberto e Celma, meu irmão Átila, e aos meus amigos próximos e da universidade, pela compreensão, incentivo e apoio total ao longo da minha jornada acadêmica. Vocês foram os pilares que me sustentaram e me fizeram persistir a cada passo desta minha jornada.

Aos meus honrados professores, quero manifestar minha profunda gratidão por todos os conselhos, compromissos e conhecimentos que compartilharam em sala de aula e comigo ao longo desta jornada acadêmica. Sobretudo, ao meu orientador Luiz Henrique, pela paciência, apoio e motivação durante todo esse período. Seus ensinamentos e conselhos foram essenciais para o meu crescimento intelectual e minha formação.

Dedico este trabalho a cada pessoa que, de alguma forma, me ajudou e contribuiu nesta minha jornada acadêmica. Tenho certeza de que, sem vocês, não conseguiria. A todos vocês, minha sincera gratidão.

“O Dado mais importante que separa o ser humano
de todos os seus irmãos e primos da escala filogenética
é o conhecimento; só o conhecimento liberta o homem, só através
do conhecimento o homem é livre e em sendo livre: ele pode aspirar
uma condição melhor de vida para ele e todos os seus semelhantes.
(Dr° Enéas Ferreira Carneiro)

Resumo

O Observatório do Mercado Imobiliário de Aracaju (OMI-AJU), vinculado à Secretaria Municipal da Fazenda de Aracaju (SEMFAZ-AJU), dedica-se ao levantamento e à análise de dados do mercado imobiliário de Aracaju. Um dos objetivos do OMI-AJU é estabelecer modelos de regressão que permitam explicar e predizer o preço de um imóvel, dadas algumas das suas características, como, por exemplo, o bairro onde ele está localizado, o nível de infraestrutura do local, sua idade, a quantidade de vagas de estacionamento que ele possui, etc. Com a utilização desses dados, foi realizada uma análise e desenvolvido um modelo de mistura de regressão linear, utilizando o algoritmo EM e o critério Bayesiano (BIC) para selecionar o número de grupos (modelos) com base nas características dos imóveis. Após as análises, percebeu-se que o maior valor de um imóvel em Aracaju, com base nas informações disponíveis, foi de R$ 39.000.000. Esse valor foi dividido em três tipos de registro: oferta, transação e ITBI, sendo o ITBI o que continha mais observações, com cerca de 36,41%. Quanto à técnica de mistura de modelos de regressão, utilizando três repetições e um número máximo de 10 grupos, através do critério BIC, foram definidos 5 grupos (modelos). O grupo 5 foi o que apresentou a maior probabilidade de uma observação pertencer a ele, com pouco mais de 34%, enquanto o grupo 3 teve a menor probabilidade, cerca de 2,77%. Através do gráfico de dispersão dos grupos, utilizando latitude e longitude, observou-se que a distribuição dos grupos é bem espalhada, indicando uma alta variabilidade entre os grupos.

Palavras-chave: Mercado imobiliário. Modelo de regressão. Mistura de modelos. Critério Bayesiano (BIC).

Abstract

The Real Estate Market Observatory of Aracaju (OMI-AJU), linked to the Municipal Finance Department of Aracaju (SEMFAZ-AJU), is dedicated to collecting and analyzing data on the real estate market in Aracaju. One of the objectives of OMI-AJU is to establish regression models that allow explaining and predicting the price of a property, given some of its characteristics, such as, for example, the neighborhood where it is located, the level of infrastructure of the place, its age, the number of parking spaces it has, etc. Using this data, an analysis was performed and a linear regression mixture model was developed, using the EM algorithm and the Bayesian Criterion (BIC) to select the number of groups (models) based on the characteristics of the properties. After the analyses, it was found that the highest value of a property in Aracaju, based on the information available, was R$ 39,000,000. This value was divided into three types of record: offer, transaction and ITBI, with ITBI containing the most observations, with approximately 36.41%. Regarding the regression model mixing technique, using three repetitions and a maximum number of 10 groups, through the BIC criterion, 5 groups (models) were defined. Group 5 was the one with the highest probability of an observation belonging to it, with just over 34%, while group 3 had the lowest probability, approximately 2.77%. Through the scatter plot of the groups, using latitude and longitude, it was observed that the distribution of the groups is well spread, indicating a high variability between the groups.

Keywords: Real estate market. Regression model. Mixture of models. Bayesian Criterion (BIC).

Resumen

El Observatorio del Mercado Inmobiliario de Aracaju (OMI-AJU), vinculado a la Secretaría Municipal de Finanzas de Aracaju (SEMFAZ-AJU), se dedica a recopilar y analizar datos sobre el mercado inmobiliario de Aracaju. Uno de los objetivos del OMI-AJU es establecer modelos de regresión que permitan explicar y predecir el precio de un inmueble, dadas algunas de sus características, como, por ejemplo, el barrio donde se ubica, el nivel de infraestructura del lugar, su antigüedad, el número de plazas de aparcamiento con las que cuenta, etc. Con estos datos se realizó un análisis y se desarrolló un modelo de regresión lineal mixta, utilizando el algoritmo EM y el criterio bayesiano (BIC) para seleccionar el número de grupos (modelos) en función de las características de las propiedades. Tras el análisis, se constató que el valor más alto de un inmueble en Aracaju, según la información disponible, fue de R$ 39.000.000. Este valor se dividió en tres tipos de registro: oferta, transacción e ITBI, siendo el ITBI el que contiene la mayor cantidad de observaciones, con aproximadamente 36,41 %. Respecto a la técnica de mezcla de modelos de regresión, utilizando tres repeticiones y un número máximo de 10 grupos, a través del criterio BIC, se definieron 5 grupos (modelos). El grupo 5 tuvo la mayor probabilidad de que una observación perteneciera a él, con poco más del 34 %, mientras que el grupo 3 tuvo la probabilidad más baja, alrededor del 2,77 %. A través del diagrama de dispersión de los grupos, utilizando latitud y longitud, se observó que la distribución de los grupos está bien dispersa, lo que indica una alta variabilidad entre los grupos.

Palabras clave: Mercado inmobiliario. Modelo de regresión. Mezcla de modelos. Criterio bayesiano (BIC).

Lista de ilustrações

Lista de tabelas

Sumário

1 INTRODUÇÃO

Análise de regressão se refere à aplicação de técnicas estatísticas, com o objetivo de identificar e modelar alguma associação entre uma variável, denominada resposta, e um conjunto de outras variáveis, denominadas preditoras. Tal associação pode ser utilizada, por exemplo, para predizer, a resposta em termos das preditoras (GRAYBILL; IYER, 1994, p. 73).

Pode-se dizer que um modelo de regressão consiste numa equação, a qual relaciona a variável resposta a uma ou mais variáveis preditoras e por suposições sobre essa equação e sobre os termos que a compõem.

Um modelo de regressão comumente utilizado é o modelo linear normal, o qual pode ser formulado por (NETER; WASSERMAN; KUTNER, 1983, p. 230):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

onde Y é a resposta, $x_1, x_2, \cdots, x_p$ são as preditoras, ϵ é uma variável aleatória, denominada erro aleatório, tal que $\epsilon \sim N(0, \sigma^2)$ e $\beta_0, \beta_1, \cdots, \beta_p$ são os coeficientes de regressão.

No modelo linear normal, considera-se que as preditoras são conhecidas. Se Y e $x_1, \cdots, x_p$ estão associadas segundo um modelo linear normal, então Y é uma variável aleatória, a qual segue distribuição normal, com

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad \text{e} \quad Var(Y) = \sigma^2.$$

Ou seja, de acordo com o modelo de regressão linear normal,

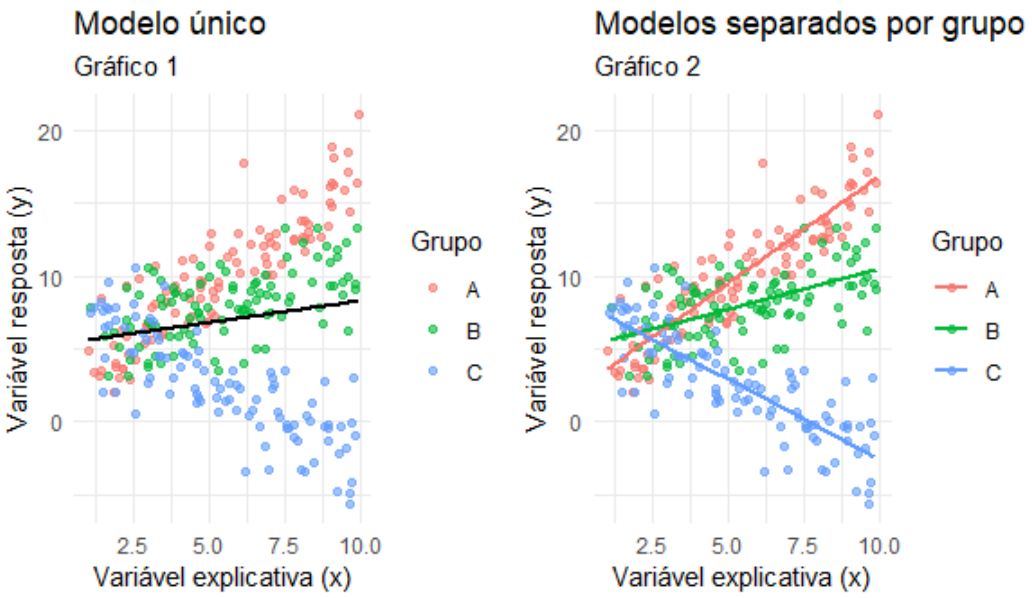
1. a resposta varia, em média, linearmente com as variáveis preditoras, sendo β_j a taxa com a qual a média de Y varia em relação a x_j, para cada $j \in \{1, \cdots, p\}$;
2. β_0 é o valor médio da resposta quando todas as preditoras são zero;
3. a variância da resposta é constante e igual à variância do erro aleatório.

Os coeficientes de regressão $\beta_0, \beta_1, \cdots, \beta_p$ e a variância do erro aleatório σ^2 são os parâmetros do modelo linear normal. Dada uma amostra da resposta e os respeoctivos valores das preditoras, o objetivo é estimar os parâmetros do modelo linear normal. Em geral, soupõe-se que a amostra selecionada é uma amostra aleatória simples e o método de estimação padrão é o método dos mínimos quadrados.

O modelo linear normal assume que os seus parâmetros são constantes ao longo da população por ele representada. Entretanto, em várias circunstâncias, tal pressuposto não é plausível. Alguns fatores que implicam isso são a heterogeneidade da população, como, por exemplo, as variáveis terem uma relação de variabilidade dependendo de subgrupos dessa mesma população. Alguns exemplos que podem ocorrer são: a diferença entre homens e mulheres (em um estudo específico), diferenças sociais e socioeconômicas, físicas, etc.

Na Figura 1, é apresentado um gráfico de dispersão de dados simulados, considerando as variáveis x, y e os grupos.

Figura 1 – Gráfico de dispersão



Fonte: Elaboração própria

Pode-se observar, pelo Gráfico 1, que, se fosse traçada uma única reta de regressão, ela não refletiria corretamente a relação entre x e y dentro de cada grupo. Como consequência, o modelo não conseguiria captar adequadamente a variabilidade dos dados, subestimando ou superestimando a relação entre essas variáveis, o que indica que essa relação não é homogênea na população total. Já no Gráfico 2, onde é traçada uma reta de regressão para cada grupo, consegue-se captar a variabilidade dos dados com mais precisão. Dessa forma, verifica-se que não é plausível assumir que os coeficientes de regressão são constantes ao longo da população, pois há diferenças significativas entre os grupos "A", "B" e "C", evidenciando a heterogeneidade na relação entre x e y.

Uma forma de modelar a heterogeneidade presente nos dados, porém não passível de ser observada, é por meio de uma mistura finita de modelos de regressão (GRÜN; LEISCH, 2007).

Uma mistura finita de modelos lineares normais assume que a população é composta por um número K de subpopulações, cada uma descrita por um modelo de regressão linear normal distinto. A mistura finita assume ainda que a subpopulação à qual uma observação pertence é desconhecida.

Para cada $k \in \{1, \cdots, K\}$, seja π_k a probabildade de uma observação pertencer à subpopulação k. Uma mistura de modelos de regressão lineares normais pode ser formulada da seguinte maneira (FARIA; SOROMENHO, 2010):

$$Y = \begin{cases} \mathbf{x}^T\boldsymbol{\beta}_1 + \epsilon_1, & \text{com probabilidade } \pi_1, \\ \mathbf{x}^T\boldsymbol{\beta}_2 + \epsilon_2, & \text{com probabilidade } \pi_2, \\ \vdots & \\ \mathbf{x}^T\boldsymbol{\beta}_K + \epsilon_K, & \text{com probabilidade } \pi_J. \end{cases} \tag{1.1}$$

onde

1. Y é a resposta;
2. $\mathbf{x}^T = (1, x_1, \cdots, x_p)$ é o vetor $(p+1)$-dimensional das preditoras com o valor 1 na primeira coordenada, representando o intercepto;
3. $\boldsymbol{\beta}_k^T = (\beta_{k0}, \cdots \beta_{kp})$, com $k \in \{1, \cdots, K\}$, denota o vetor $(p+1)-$dimensional dos coeficientes da regressão para o $j-$ésimo componente;
4. para cada $k \in \{1, \cdots, K\}$, π_k é a probabilidade de uma observação pertencer à subpopulação k;
5. $\epsilon_1, \cdots, \epsilon_K$ são os erros aleatórios, os quais são variáveis aleatórias tais que $\epsilon_k \sim N(0, \sigma_k^2)$ para cada $k \in \{1, \ldots, K\}$.

Vê-se que $\pi_1, \cdots, \pi_K$ satisfazem $0 < \pi_k < 1$ e $\sum_j^K \pi_k = 1$.

Uma mistura de modelos de regressão pode ser utilizada para representar a presença de subpopulações dentro de uma população geral, sem exigir que um conjunto de dados observados identifique explicitamente a subpopulação à qual uma observação individual pertence. Assim, a mistura oferece uma abordagem coerente para a situação mencionada anteriormente, na qual os parâmetros do modelo de regressão, tanto os coeficientes quanto a variância do erro, variam ao longo da população.

Misturas de modelos lineares normais são flexíveis o suficiente para capturar estruturas complexas de variabilidade na amostra e, ao mesmo tempo, interpretáveis. De fato, cada componente de uma mistura é um modelo linear normal e, como tal, sua interpretação pode ser feita da maneira convencional.

O observatório do Mercado Imobiliário de Aracaju (OMI-AJU), vinculado à Secretaria Municipal da Fazenda de Aracaju (SEMFAZ-AJU), dedica-se ao levantamento e à analise de dados do mercado imobiliário de Aracaju. Um dos objetivos do OMI-AJU é estabelecer modelos de regressão que permitam explicar e predizer o preço de um imóvel, dadas algumas das suas caraterísticas como, por exemplo, bairro onde ele está localizado, nível de infraestrutura do local, sua idade, quantidade de vagas de estacionamento que ele possui, etc. Os modelos obtidos são utilizados para estimar preços dos imóveis e tais estimativas são utilizadas no cálculo de impostos como, por exemplo, o ITBI. Para esse fim, o observatório utiliza modelagem de regressão e o principal modelo adotado vem sendo o modelo de linear normal.

O presente trabalho visa contribuir com o OMI-AJU propondo o uso de misturas de modelos de lineares normais para modelar o preço (resposta) dos imóveis residenciais no município de Aracaju-SE em termos das características desses imóveis. A adoção de um modelo baseado em uma mistura é justificada pelo fato de haver a necessidade, por parte do OMI-AJU, de um modelo flexível, porém interpretável, características presentes na mistura de modelos de regressão lineares normais.

O presente trabalho está dividido em 6 tópicos, iniciando pelo Capítulo 1, a Introdução, onde é apresentado de maneira breve o conteúdo abordado ao longo do estudo. No segundo capítulo, são tratados os objetivos a serem analisados. Já no Capítulo 3, é apresentada a justificativa deste trabalho. No Capítulo 4, é apresentada a revisão de literatura, que aborda o mercado imobiliário, a precificação de imóveis e os métodos para a predição de imóveis, através de regressão linear, machine learning e os métodos utilizados nesse processo, que envolvem misturas de modelos. No Capítulo 5, é abordada a metodologia aplicada neste estudo, que também envolve misturas de modelos. São discutidos a estimativa dos parâmetros pelo algoritmo EM, o critério para seleção de grupos, a plataforma computacional e o conjunto de dados. O Capítulo 6 trata-se dos resultados obtidos, através da análise descritiva, seguida pela aplicação da técnica de misturas de modelos de regressão linear para a construção. Inicia-se com a apresentação de algumas variáveis dos dados e a técnica de misturas de modelos, incluindo o machine learning, as árvores de decisão e o random forest. O penúltimo capítulo apresenta os resultados. E, por fim, no Capítulo 7, o estudo é finalizado com a apresentação das conclusões e possíveis trabalhos futuros que possam ser realizados a partir deste trabalho.

2 OBJETIVOS

2.1 Geral

objetivo deste trabalho consiste em construir uma mistura de modelos capaz de precificar o valor dos imóveis de Aracaju - SE, utilizando o algoritmo de misturas de modelos de regressão linear.

2.2 Específicos

- Realizar o pré-processamento dos dados, verificando a duplicidade, se a variável apresenta variabilidade, não tendo comportamento constante, observar se há dados faltantes, transformar variáveis caso seja necessário, entre outras análises.
- "Ajustar uma mistura de modelos lineares, que segue uma distribuição normal padrão, e definir a quantidade de subgrupos da mistura com base no critério de informação bayesiano (BIC).
- "Interpretar os parâmetros e grupos, observando como cada componente se distribui entre os grupos, identificando quais variáveis têm maior impacto no valor do imóvel e quais não têm. Além disso, analisar a variabilidade dos subgrupos por meio do desvio padrão.

3 REVISÃO LITERÁRIA

Um fator importante no empreendedorismo imobiliário é a precificação de um imóvel, a qual pode ocorrer antes da construção (na planta) ou em transações. É necessário saber o preço do imóvel para, por exemplo, poder estimar possíveis ganhos sobre ele ou mesmo calcular impostos residenciais (ALVES et al., 2011).

A precificação de imóveis, por meio de modelos estatísticos, é bem conhecida e bastante utilizada, sendo os modelos de regressão linear, regressão não linear, lineares generalizados, redes neurais e florestas aleatórias alguns dos modelos que vêm sendo utilizados (PEREIRA; GARSON; ARAÚJO, 2012; CHAGAS, 2019). Nesse contexto, os modelos que utilizam atributos do imóvel para predizer seu preços são conhecidos como modelos hedônicos (ALVES et al., 2011). Tais modelos assumem que cada atributo que compõe o imóvel tem um preço, denominado preço sombra, e o preço do imóvel é a soma do preço sombra de cada atributo (PAIXÃO, 2015). Alguns desses atributos, como a área construída, o padrão construtivo e o número de vagas na garagem, são de natureza estrutural, enquanto outros, como o bairro onde o imóvel está situado e a distância em relação aos polos de influência, são de natureza locacional (PINTO; FERNANDES, 2019).

Alves et al. (2011) aplicaram modelagem de regressão para estimar os preços dos imóveis paulistanos e identificar de que forma as condições de financiamento e, sobretudo, o boom no preço das ações das empresas do mercado imobiliário impactaram o valor dos imóveis residenciais na cidade de São Paulo.

Nunes, Neto e Freitas (2019) utilizaram um modelo de regressão linear múltipla para a determinação do valor de mercado de apartamentos residenciais na cidade de Fortaleza, CE. Um fator interessante nesse trabalho foi a utilização da técnica de "ridge regression"para solucionar o problema de multicolinearidade. Um dos resultados obtidos foi que algumas variáveis independentes apresentaram correlação com a variável dependente. Com isso, não foram selecionadas para o modelo final, mas isso não significa que elas não sejam importantes para entender o comportamento monetário do imóvel. Outro fator interessante é que o estado (restrito somente ao local de pesquisa) poderia utilizar esse modelo para cálculos de impostos referentes ao imóvel, como, por exemplo, o ITBI. Um dos resultados obtidos é que as interações das variáveis boom com as características dos imóveis e spread (risco de crédito) mostraram que, em períodos de boom, algumas características dos imóveis perdem importância na formação dos preços, enquanto as condições de financiamento imobiliário ganham relevância.

O trabalho de Medeiros e Carvalho (2017), cujo objetivo era formular um modelo de regressão para avaliar os preços de aluguéis dos apartamentos do município de Petrópolis-RJ, apresentou os seguintes resultados: 77% da variabilidade dos preços de aluguéis de apartamentos

nesse município são explicados por 4 variáveis, sendo a variável "vaga de garagem"a que possui maior impacto no modelo. Também foi salientado que é necessário atualizar periodicamente os cálculos do modelo, pois o mercado imobiliário passa por várias dinâmicas ao longo do tempo.

No município de Aracaju-SE, alguns trabalhos encontrados que utilizam técnicas estatísticas de modelagem para modelar o preço de imóveis foram os trabalhos de Sousa (2023), Santos (2023). No trabalho de Sousa (2023), a autora utilizou um algoritmo de machine learning, o Random Forest, para desenvolver um modelo de previsão capaz de precificar imóveis com base em suas características no município de Aracaju-SE. Ela utilizou a técnica de web scraping para coletar os dados e, através disso, conseguiu imóveis que variavam de 80K (mil) até 2M (milhões). Conseguiu identificar que, na zona de expansão e na zona Sul de Aracaju, o preço médio de venda é maior do que nas demais zonas. Identificou também que suítes são um fator importante que influencia bastante o preço do imóvel, com quantidades que variam de 0 a 7 suítes por imóvel. Algo importante a se salientar é que este algoritmo de machine learning (Random Forest) é capaz de prever bem os preços dos imóveis a partir de novas informações, até 1M, sendo a variável "Área"a que possui maior importância no modelo. Já no trabalho de Santos (2023), ele utilizou regressão linear múltipla para modelar os preços dos imóveis na zona sul do município de Aracaju-SE, nos bairros de Atalaia, Aruana e Coroa do Meio, com base na distância desses imóveis à praia, para verificar se, quanto mais próximos esses imóveis estivessem da praia, eles eventualmente teriam uma valorização imobiliária. Ele conseguiu identificar que as variáveis “Setor Urbano” e “Distância à Praia” são correlacionadas, o que acaba gerando no modelo de regressão o problema de multicolinearidade e também conseguiu um modelo com um coeficiente de determinação em torno de 81%. Por fim, conseguiu concluir que os imóveis localizados nos bairros da Atalaia e Coroa do Meio sofrem uma maior influência da distância à praia.

Até o presente momento, não foram encontrados trabalhos aplicando misturas de modelos de regressão para modelar preços de imóveis no Brasil. No âmbito internacional, pode-se mencionar o trabalho de (HUANG; WANG, 2013), que utilizou misturas de modelos de regressão não paramétricos para modelar o índice HPI de preços de imóveis nos EUA em termos do PIB. Eles mostraram que o impacto do PIB, na mudança do HPI, apresenta padrões diferentes em diferentes ciclos macroeconômicos, o que fortalece a conexão com o uso das misturas. Foi identificado também que existem dois ciclos econômicos durante o período estudado, o que sugere o uso de dois modelos. No primeiro componente (Primeiro ciclo), a variável PIB tem impacto positivo na mudança do HPI, já no segundo ciclo, quando o PIB está no meio, o crescimento do HPI tende a ser menor em comparação ao alto e baixo crescimento do PIB. Porém, o resultado não foi considerado satisfatório, pois os valores do HPI e do PIB exibiram autocorrelação, enquanto o procedimento de ajuste proposto por eles pressupõe independência entre as observações.

4 METODOLOGIA

4.1 Formulação da mistura de modelos de regressão linear e interpretação da mistura

Uma mistura finita de modelos lineares normais assume que a população é composta por um número K de subpopulações, cada uma descrita por um modelo de regressão linear normal distinto. A mistura finita assume ainda que a subpopulação à qual uma observação pertence é desconhecida.

Para cada $k \in \{1, \cdots, K\}$, seja π_k a probabildade de uma observação pertencer à subpopulação k. Uma mistura de modelos de regressão lineares normais pode ser formulada por (FARIA; SOROMENHO, 2010):

$$Y = \begin{cases} \mathbf{x}^T \boldsymbol{\beta}_1 + \epsilon_1, & \text{com probabilidade } \pi_1, \\ \mathbf{x}^T \boldsymbol{\beta}_2 + \epsilon_2, & \text{com probabilidade } \pi_2, \\ \vdots & \\ \mathbf{x}^T \boldsymbol{\beta}_K + \epsilon_K, & \text{com probabilidade } \pi_K. \end{cases} \tag{4.1}$$

onde

1. Y é a resposta;
2. $\mathbf{x}^T = (1, x_1, \cdots, x_p)$ é o vetor $(p+1)$-dimensional das preditoras com o valor 1 na primeira coordenada, representando o intercepto;
3. $\boldsymbol{\beta}_k^T = (\beta_{k0}, \cdots \beta_{kp})$, com $k \in \{1, \cdots, K\}$, denota o vetor $(p+1)-$dimensional dos coeficientes da regressão para o $j-$ésimo componente;
4. para cada $k \in \{1, \cdots, K\}$, π_k é a probabilidade de uma observação pertencer à subpopulação k;
5. $\epsilon_1, \cdots, \epsilon_K$ são os erros aleatórios, os quais são variáveis aleatórias tais que $\epsilon_k \sim N(0, \sigma_k^2)$ para cada $k \in \{1, \ldots, K\}$.

Como a mistura presume que existem várias subpopulações na população, cada uma descrita por um modelo específico de regressão linear normal com uma probabilidade associada, é possível identificar essas subpopulações sem informações prévias sobre elas.

Como sabemos que, na regressão linear normal tradicional, atribui-se uma única relação entre as variáveis, a mistura de modelos de regressão linear assume que a população é composta

por diferentes subpopulações, cada uma com sua própria relação entre as variáveis. Assim, cada observação da variável resposta y_i pertence a uma determinada subpopulação com uma probabilidade π_k.

Os coeficientes β_k dessas variáveis variam entre esses grupos, de modo que, para cada grupo, esses coeficientes assumem valores distintos. Por exemplo, uma determinada variável X_j pode ter coeficiente β_k positivo para o grupo 1, enquanto para o grupo 2 esse coeficiente pode ser negativo. Isso é algo que pode ocorrer dentro da mistura de modelos, pois essa abordagem permite capturar a heterogeneidade dos dados.

Dessa forma, podemos interpretar as misturas de modelos como uma ferramenta para identificar e diferenciar essas subpopulações (grupos), destacando suas distintas relações entre as variáveis. Isso possibilita uma melhor compreensão das tendências implícitas na população e permite uma análise mais detalhada dos dados.

Dentro de cada grupo, os coeficientes de regressão podem ser interpretados da maneira convencencional. Isto é, o coeficiente β_{kj} descreve a variação da resposta média $E(Y)$ no grupo k, correspondente à preditora x_j: se $\beta_{kj} > 0$ ($\beta_{kj} < 0$), quado x_{jk} aumenta uma unidade, a resposta, no grupo k, aumenta (diminue), em média, β_{kj} unidades.

4.2 Estimação dos parâmetros pelo algoritmo EM

Seja$y_1, y_2, \ldots, y_n$ uma amostra aleatória simples da resposta Y e $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ os respectivos valores do vetor de preditoras $\mathbf{x}$. Supõe-se que a população da qual essa amostra foi selecionada segue uma mistura de modelos lineares normais, com K componentes.

Seja $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$ o vetor contendo todos os parâmetros da mistura. Aqui, adota-se o estimador de máxima verossimilhança para estimar $\boldsymbol{\theta}$, o qual corresponde ao valor de $\boldsymbol{\theta}$ que maximiza a log-verossimilhança do modelo, com base na amostra dada:

$$l(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} \pi_j \phi_j(y_i|\mathbf{x}_i)\right), \tag{4.2}$$

No contexto das misturas de modelos lineares normais, o método padrão para o cálculo da estimativa de máxima verossimilhança do vetor de parâmetros do modelo é o algoritmo EM (FARIA; SOROMENHO, 2010).

O algoritmo EM é uma abordagem amplamente aplicável para o cálculo iterativo de estimativas de máxima verossimilhança quando as observações podem ser vistas como incompletas (MCLACHLAN; KRISHNAN, 2008, p. 1). A ideia aqui é considerar os dados como compostos por trios (x_i, Y_i, Z_i), onde Z_i ie uma variável aleatória, não observada, que especifica o componente da mistura do qual a observação y_i foi extraída. Nesse caso, Z_i é a a porção dos

dados que não é observada, isto é, que torna os dados incompletos.

O algoritmo EM procede iterativamente em duas etapas: **E** (cálculo da esperança condicional da log-verossimilhança) e **M** (de Maximização da esperança condicional da log-verossimilhança).

Seja $\boldsymbol{\theta}^{(r)}$ a estimativa dos parâmetros após a r-ésima iteração. Na $(r+1)$-ésima iteração, a etapa E do algoritmo EM envolve o cálculo da função Q, que é a esperança da log-verossimilhança dos dados completos, condicionalmente às estimativas atuais dos parâmetros e aos dados observados:

$$Q(\theta, \theta^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}^{(r)} \phi_j(y_i|x_i), \tag{4.3}$$

onde ϕ_k denota a função de densidade de probabilidade de Y correspodnente ao grupo k, e

$$w_{ik}^{(r)} = \frac{\pi_k^{(r)} \phi_j(y_i|x_i)}{\sum_{k=1}^{K} \pi_k^{(r)} \phi_k(y_i|x_i)} \quad (i = 1, ..., n; k = 1, ..., K) \tag{4.4}$$

é a estimativa da probabilidade *a posteriori* de que a i-ésima observação pertença ao j-ésimo componente da mistura.

A etapa M calcula uma estimativa atualizada $\theta^{(r+1)}$, maximizando a função Q em relação a θ. Isso equivale a calcular a proporção amostral e as estimativas de mínimos quadrados ponderados ao realizar uma regressão ponderada de $y_1, ..., y_n$ sobre $x_1, ..., x_n$ com pesos $w_{1k}, ..., w_{nk}$ $(k = 1, ..., K)$.

Dessa forma, na etapa M da $(r+1)$-ésima iteração, o ajuste atual dos parâmetros é dado explicitamente por:

$$\hat{\pi}_j^{(r+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(r)}}{n} \quad (k = 1, ..., K) \tag{4.5}$$

$$\hat{\beta}_k^{(r+1)} = (X^T W_k X)^{-1} X^T W_k Y \quad (k = 1, ..., K), \tag{4.6}$$

onde X é uma matriz $n \times (p+1)$ de preditores, W_k é uma matriz diagonal $n \times n$ com entradas diagonais $w_{ik}^{(r)}$ e Y é um vetor $n \times 1$ da variável resposta.

Além disso,

$$\hat{\sigma}_k^{2(r+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(r)} (y_i - x_i^T \hat{\beta}_k^{(r+1)})^2}{\sum_{i=1}^{n} w_{ik}^{(r)}} \quad (k = 1, ..., K). \tag{4.7}$$

As etapas **E** e **M** são alternadas repetidamente até q haja convergência do valor da log-verossimilhança dada por 4.2.

4.3 Seleção do número de grupos mais adequado com base no BIC

Para selecionar o número de grupos gerados pelo algoritmo EM, utilizou-se o critério de informação Bayesiano (BIC) (MCLACHLAN; PEEL, 2004, p. 175, 209). O BIC é definido como:

$$BIC = 2\ln(\hat{L}) - k\ln(n), \tag{4.8}$$

onde $\hat{L}$ é o estimador de máxima verossimilhança de L, k é o número de parâmetros a ser estimado no modelo e n é o tamanho da amostra.

Na estatística, o BIC adiciona o termo $-k\ln(n)$ à função de log-verossimilhança para penalizar a complexidade do modelo com o aumento do número de componentes. Ajustou-se um modelo para cada quantidade de grupos fixada, identificando e selecionando aquele que apresentou o menor BIC. Foi utilizada uma estratégia em que cada ajuste gerado contou com três repetições com valores iniciais distintos. Dentre esses ajustes, foi escolhido aquele que apresentou a maior máxima verossimilhança.

4.4 Conjunto de dados e pré-processamento

O observatório do Mercado Imobiliário de Aracaju (OMI-AJU), vinculado à Secretaria Municipal da Fazenda de Aracaju (SEMFAZ-AJU), dedica-se ao levantamento e à analise de dados do mercado imobiliário de Aracaju. Um dos objetivos do OMI-AJU é estabelecer modelos de regressão que permitam explicar e predizer o preço de um imóvel, dadas algumas das suas caraterísticas como, por exemplo, bairro onde ele está localizado, nível de infraestrutura do local, sua idade, quantidade de vagas de estacionamento que ele possui, etc. Os modelos obtidos são utilizados para estimar preços dos imóveis e tais estimativas são utilizadas no cálculo de impostos como, por exemplo, o ITBI. Para esse fim, o observatório utiliza modelagem de regressão e o principal modelo adotado vem sendo o modelo de linear normal.

O conjunto de dados, formado por variáveis, as quais representam características de imóveis, localizados no município de Aracaju, foi fornecido pelo observatório do Mercado Imobiliário de Aracaju (OMI-AJU), vinculado à Secretaria Municipal da Fazenda de Aracaju (SEMFAZ-AJU). O conjunto de dados contém 9.918 observações e 41 variáveis. Dentre essas variáveis, havia uma no formato de caractere e 40 numéricas.

Primeiramente, foi realizada uma reunião com os técnicos da SEMFAZ para entender o que cada variável representava no conjunto de dados e quais seriam mais importantes, teoricamente, para a predição do preço do imóvel.

Após receber o conjunto de dados, foi criada a variável `tipo_registro_valor`, que indica se o valor do imóvel se refere a uma transação, oferta ou ITBI. Notamos a existência de duplicidades nos dados, pois alguns imóveis possuíam registros para oferta, ITBI e transação. Para evitar redundâncias, consideramos apenas o valor mais recente do imóvel, seguindo a ordem de prioridade: transação − ITBI − oferta.

Após essa etapa, removemos as variáveis previamente descartadas na reunião com os técnicos. As variáveis que permaneceram passaram por transformações para garantir que estivessem no formato adequado: as qualitativas foram convertidas em fatores e as quantitativas em valores numéricos. Para essas variáveis, analisamos a variabilidade individual a fim de verificar se deveriam ser incluídas no modelo. Variáveis preditoras com baixa variabilidade foram descartadas, tanto qualitativas quanto quantitativas.

Algumas variáveis qualitativas, como `conservacao`, `ind_loc_peso` e `numpav`, foram recategorizadas para aumentar sua variabilidade. A variável `anoidade` (ano de construção do imóvel), que originalmente era numérica, foi transformada em categórica para facilitar a interpretação dos resultados.

Em seguida, analisamos a presença de valores faltantes no conjunto de dados e identificamos 31 observações ausentes na variável `pontuacao`. Em sua escala original, essa variável apresentava correlação de 0,15 com o valor do imóvel (variável resposta), enquanto na escala logarítmica a correlação aumentava para 0,45. No entanto, nenhuma variável preditora, em qualquer escala, apresentou correlação significativa com `pontuacao`. Por isso, optamos, a princípio, por removê-la. Também removemos as variáveis `vlr_area` e `vlr_unit`, pois eram derivadas da variável resposta `valor`.

Posteriormente, analisamos a correlação entre as variáveis preditoras para evitar multicolinearidade no modelo e, ao mesmo tempo, realizar uma pré-seleção das variáveis mais relevantes. Como exemplo, removemos as variáveis `cota_terreno`, `prof_equiv` e `tesprin`, pois estavam representadas pela variável `stlote`. Além disso, realizamos testes de transformação das variáveis quantitativas para verificar quais formatos apresentavam melhor correlação com a variável resposta. Algumas variáveis foram transformadas utilizando logaritmo, enquanto outras foram ajustadas com raiz quadrada.

Por fim, após todas as transformações feitas, o conjunto de dados final ficou com 7.770 observações e 17 variáveis, sendo 8 delas do tipo `dummy` (fatores) e 9 numéricas.

4.5 Software R

Os cálculos foram feitos por meio de rotinas computacionais implementadas na ferramenta RStudio, que é a plataforma de hospedagem do software R 4.3.3 (R Core Team, 2024). O pacote `flexmix` contém as funções necessárias para a modelagem por mistura de modelos de regressão (LEISCH, 2004; GRUEN; LEISCH, 2025).

O código desenvolvido encontra-se no apêndice script.

5 RESULTADOS

5.1 Análise Descritiva

Na tabela 1, é apresentada uma descrição das variáveis finais que ficaram no conjunto de dados, incluindo o nome de cada uma, o que cada uma representa e se ela é categórica (fatorial) ou numérica.

Tabela 1 – Descrição e tipos das variáveis presente no conjunto de dados.

Variável	**Descrição**	**Tipo**
av_rod	Indica se o imóvel está localizado em avenida ou rodovia (0 – Não; 1 – Sim)	Categórica (Factor)
conservacao	Estado de conservação do imóvel (1 – Não conservado; 2 – Desgaste; 3 – Conservado)	Categórica (Factor)
coordx	Longitude em UTM do centróide do lote	Numérica
coordy	Latitude em UTM do centróide do lote	Numérica
densidade_vertical	Densidade de prédios verticais residenciais com mais de 4 pavimentos, representada como uma variável proxy obtida por mapa de calor.	Numérica
dist_av	Distância até a avenida ou rua de grande movimento mais próxima	Numérica
idade	Idade da construção em anos	Numérica
idh	Índice de Desenvolvimento Humano, calculado com base em dados do censo de 2010 do IBGE	Numérica
ind_loc_peso	Índice de Localização Peso (uma variável proxy para a localização do imóvel na planta de 2014)	Categórica (Factor)
infra_aju	Soma dos elementos de infraestrutura presentes na face da quadra (1 – menor ou igual a 3; 2 – igual a 4; 3 – igual a 5; 4 – igual a 6)	Categórica (Factor)
numpav	Número de pavimentos da construção no lote (1-2 ou 2-4)	Categórica (Factor)
padrao	Padrão da construção (ex: 0,5 – 0,7; 1 – 1,2 – 1,35)	Categórica (Factor)
peso	Representa o eixo de logradouro (1 – local; 2 – secundária; 3 – principal)	Categórica (Factor)
stcunid	Área total construída da unidade em metros quadrados	Numérica
stlote	Área total do lote em metros quadrados	Numérica
tipo_registro_valor	Tipo de registro do valor (ITBI, Oferta, ou Transação)	Categórica (Factor)
valor	Valor atribuído pela Prefeitura, valor da oferta ou preço da transação	Numérica

Fonte: Elaboração própria.

Na Tabela 2, é apresentada uma descritiva de mínimo, máximo, média, mediana, desvio padrão, primeiro quartil e terceiro quartil da variável resposta, que é o valor do imóvel. Podemos observar que o valor mínimo do imóvel é de apenas R$ 2.651,00, algo bastante incomum de se ter, pois é quase impossível encontrar um imóvel nesse valor. Provavelmente se deve a um erro de registro no momento do cadastramento desse imóvel. O preço médio dos imóveis é de R$

270.000,00, enquanto a mediana é de R$ 373.906,00. Observamos que esses valores estão um pouco distantes, o que sugere que a distribuição desses dados é assimétrica. O valor máximo registrado do imóvel foi de R$ 39.500.000,00. Podemos notar que o valor máximo e mínimo estão bem distantes, o que pode indicar uma alta variabilidade do conjunto de dados, mas podemos medir isso através do desvio padrão, que foi de R$ 764.347,00. Isso significa que, em média, os preços dos imóveis estão distantes da média por cerca de R$ 764.347,00. Esse alto valor do desvio padrão indica uma alta variação do preço dos imóveis.

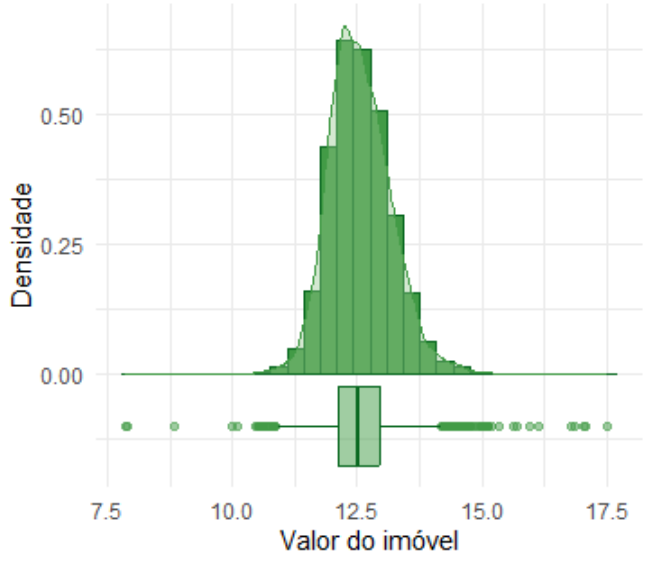
Tabela 2 – Descritiva da variável valor do imóvel.

Variável	**Mínimo**	**Média**	**Mediana**	**Máximo**	**Desvio Padrão**
Valor	2651	373906	270000	39500000	764347

Fonte: Elaboração própria.

Na Figura 2, apresenta-se o gráfico conjunto contendo o histograma, densidade e boxplot da variável 'valor do imóvel'. Porém, tivemos que transformá-la para a escala logarítmica, pois, como vimos na Tabela 2, o valor do imóvel apresenta uma alta variabilidade. Podemos observar, através dos gráficos de histograma com densidade, que os dados se aproximam de uma distribuição normal, pois apresentam um formato mais simétrico. Uma forma de corroborar com esse pensamento é através do boxplot presente, pois o valor da mediana do gráfico está praticamente no centro do histograma, e a sua diferença interquartil (tamanho da caixa) é pequena. Notamos também a presença de alguns outliers, tanto altos demais quanto baixos demais. Esses valores acabam puxando a média para baixo ou para cima.

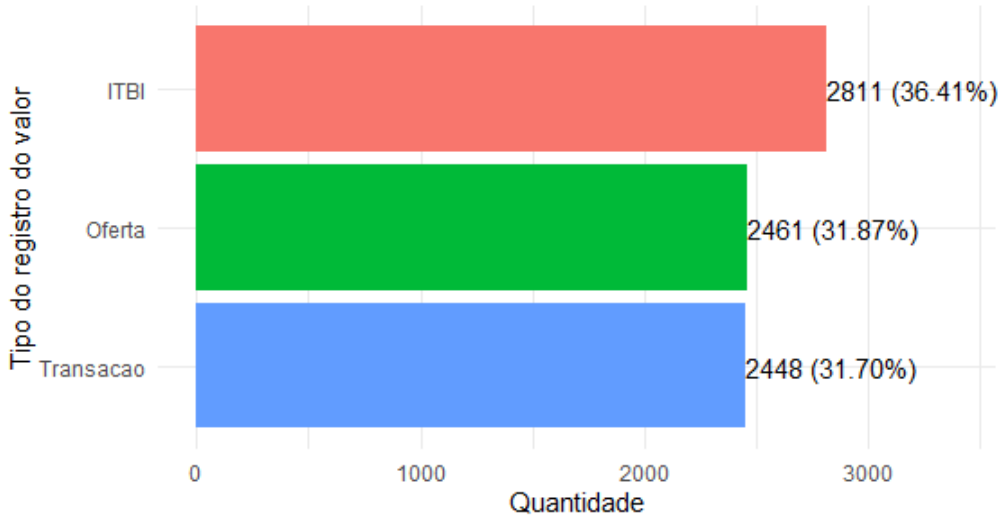
Figura 2 – Gráficos de histograma, densidade e boxplot do valor do imóvel



Fonte: Elaboração própria

Na Figura 3, apresenta-se o gráfico de barras para o tipo de registro do valor do imóvel. Podemos observar que o tipo de registro ITBI foi o que apresentou a maior quantidade, com 2.811 observações (cerca de 36,41% dos dados). Em seguida, vem o registro de Oferta, com 2.461 observações (aproximadamente 31,87% dos dados). O registro de Transação foi o que apresentou a menor quantidade, com 2.448 observações (cerca de 31,70% dos dados). De modo geral, podemos observar que os dados estão bem distribuídos entre os três tipos de registro do valor do imóvel.

Figura 3 – Gráfico de barras do tipo de registro do valor do imóvel



Fonte: Elaboração própria

Já na Tabela 3, é apresentada uma análise do preço do imóvel por tipo de registro. Fazendo uma análise separada por tipo de registro, podemos observar que, no registro por ITBI, o valor mínimo observado foi de R$ 36.500,00, enquanto o valor máximo foi de R$ 19.300.000,00, uma diferença bem alta, o que corrobora para o alto valor do desvio padrão, que foi de R$ 399.547,00. A média e a mediana foram R$ 271.198,00 e R$ 215.000,00, respectivamente. Podemos observar que os valores ainda estão distantes, mas são mais próximos do que quando se analisa o valor de modo geral. Já por oferta, é onde se apresenta a maior disparidade dos valores, sendo o valor mínimo de R$ 2.651,00 e o máximo de R$ 39.500.000,00. Esse é o tipo de registro onde há maior diferença entre os valores mínimo e máximo, bem como entre a média e a mediana. Também foi onde se apresentou o maior desvio padrão, que foi de R$ 1.241.487,00, sugerindo fortemente a presença de outliers, tanto altos quanto baixos. A média e a mediana foram, respectivamente, R$ 561.832,00 e R$ 400.000,00. Por fim, por transação, foi onde se apresentou a menor variabilidade dos valores dos imóveis, com um desvio padrão de R$ 239.692,00. Embora ainda seja um valor alto, é menor do que os demais e também quando se analisa o valor de modo geral. Os valores de mínimo e máximo foram R$ 40.000,00 e R$ 3.700.000,00, respectivamente, sendo o valor máximo o menor apresentado entre os tipos de registros. A média e a mediana foram,

respectivamente, R$ 302.920,00 e R$ 250.000,00, sendo o tipo de registro que apresentou a menor diferença entre a média e a mediana. Algo curioso a se observar é que a média é maior que a mediana em todos os tipos de registros, indicando que a maioria dos valores está abaixo da média e a presença de outliers que puxam essa média para cima.

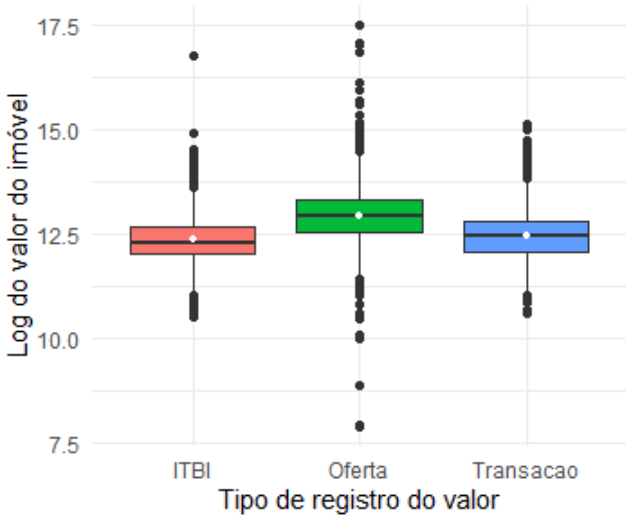
Tabela 3 – Descritiva da variável tipo de registro do valor.

Tipo de Registro	**Mínimo**	**Mediana**	**Média**	**Máximo**	**Desvio Padrão**
ITBI	36500	215000	271198	19300000	399547
Oferta	2651	400000	561832	39500000	1241487
Transação	40000	250000	302920	3700000	239692

Fonte: Elaboração própria.

Na figura 4, podemos observar o gráfico de boxplot do preço do imóvel por tipo de registro. É perceptível que a variação do valor do imóvel registrado por Oferta é superior às demais formas de registro. O ponto branco sob a linha que representa a mediana é a média, e notamos que, em todos os casos, esse ponto está um pouco localizado acima da linha (mediana), sendo o registro por ITBI o qual esse ponto está mais distante da linha. Conseguimos perceber também que, em todos os tipos de registro de valor, existem outliers, tanto extremos superiores quanto inferiores, sendo o registro por Oferta o qual apresenta os outliers mais distantes registrados.

Figura 4 – Gráfico de boxplot do tipo de registro do valor do imóvel

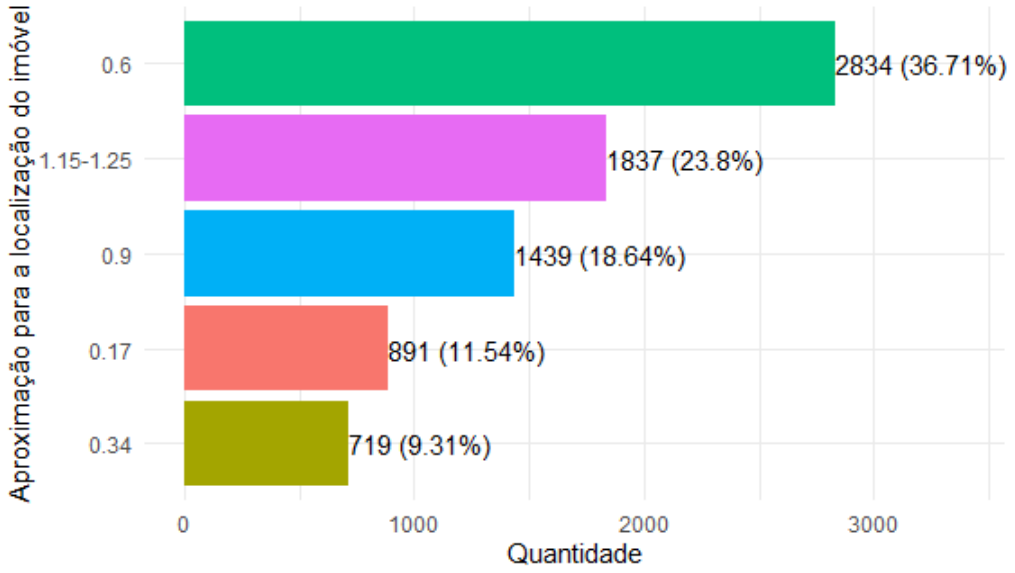


Fonte: Elaboração própria

Na Figura 5, é apresentado o gráfico de barras para a variável "ind loc peso", que representa uma aproximação para a localização do imóvel com base na planta de 2014. Podemos observar que a métrica 0.6 representa 36,71% dos dados, seguida pelas métricas 1.15-1.25 e 0.9, que representam 23,8% e 18,64% dos dados, respectivamente. A métrica de localização

que tem menos observações é a 0.34, que representa apenas 9,31% dos dados. De modo geral, podemos observar que, dentro dessa variável, os dados estão relativamente distribuídos entre os cincos tipos de aproximação para a localização do imóvel, sendo a métrica 0.6 a que contêm mais observações e a 0.34 a que contêm menos.

Figura 5 – Gráfico de barras da aproximação do imóvel



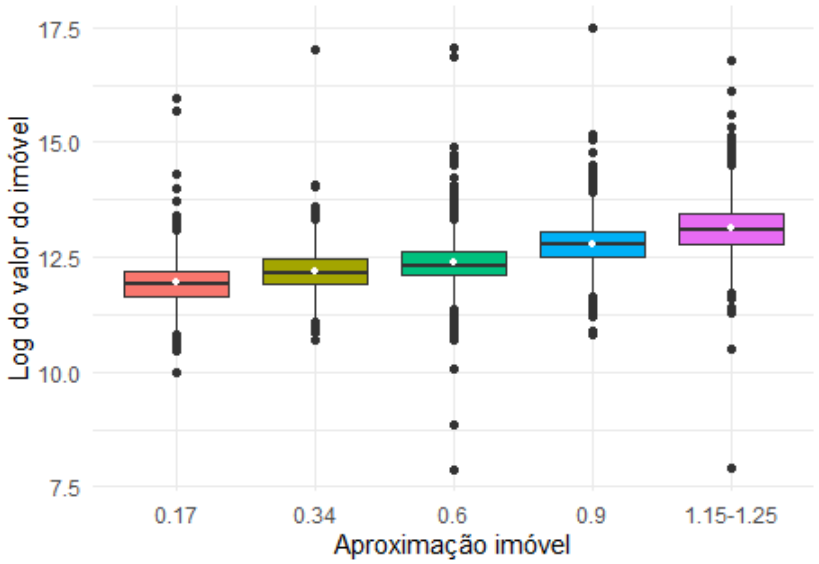
Fonte: Elaboração própria

Na figura 6, apresenta-se o gráfico de boxplot para a variável "ind loc peso". Podemos observar que a métrica 1.15-1.25 foi a que apresentou maior variação do valor do imóvel, pois é a que possui o intervalo interquartil maior. Já a métrica de aproximação que apresentou menor variação do valor do imóvel foi a 0.6, pois sua caixa é a menor. Algo curioso a se observar é que, quanto maior a métrica, maior o valor do imóvel que ela está variando. Percebemos também que, para todas as métricas, o valor da média (ponto branco) está localizado acima da mediana (linha da caixa), indicando que os outliers altos estão puxando a média para cima. A média que mais se aproximou da mediana foi a da métrica 1.15-1.25, em que o ponto está quase sobreposto à linha.

Na figura 7, é apresentado o gráfico de barras da variável "infra aju", que representa a soma dos elementos de infraestrutura presentes na face da quadra onde o imóvel está localizado. Esses elementos incluem água, esgoto, galeria pluvial, guias e sarjetas, iluminação pública e pavimentação, e são representados por: 1 – soma menor ou igual a três; 2 – soma igual a quatro; 3 – soma igual a cinco; 4 – soma igual a seis. Podemos perceber que 65,48% dos dados têm uma soma igual a seis, das características citadas, seguido pelos imóveis que têm soma igual a cinco, representando 14,3% dos dados. A soma igual a quatro contém menos observações, cerca de 7,67%.

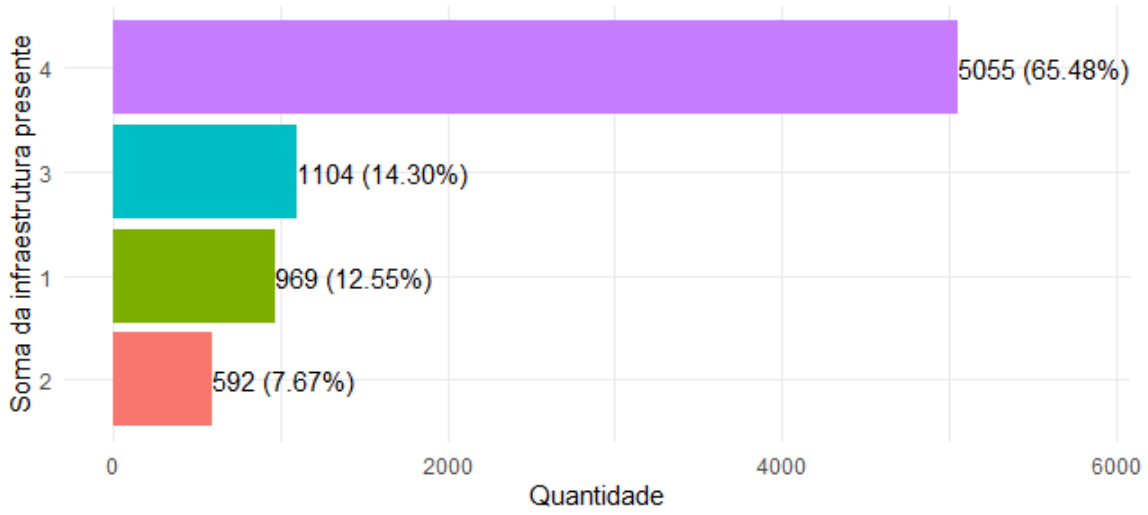
Na figura 8, é apresentado o gráfico de boxplot para a variável "infra aju". Podemos

Figura 6 – Gráfico de boxplot da aproximação do imóvel



Fonte: Elaboração própria

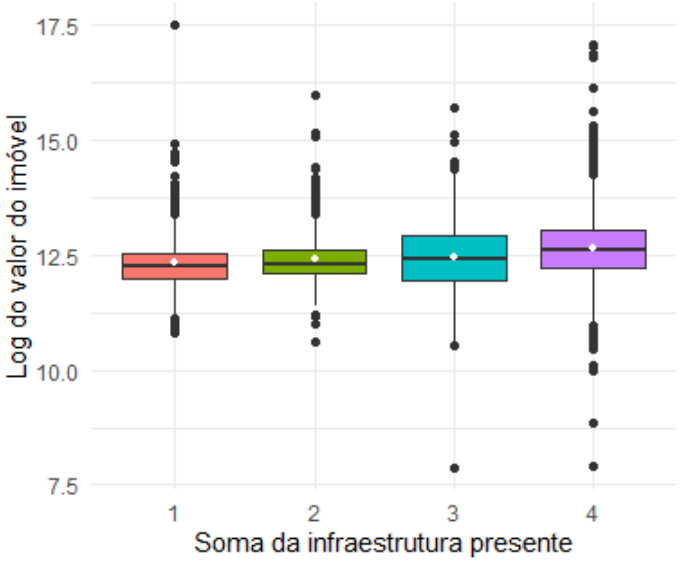
Figura 7 – Gráfico de barras da infraestrutura da quadra do imóvel



Fonte: Elaboração própria

perceber que o grupo 3, cuja soma é igual a cinco, é o que tem maior variação dentre os grupos, pois sua caixa (diferença interquartil) é a maior, e a que apresenta a menor variação é o grupo 2, sendo também o grupo 1 com a caixa bem próxima. O grupo 2 é o que sofre mais com a presença de outliers, pois é o qual a média está mais distante da mediana (o ponto branco que representa a média está mais distante da linha da caixa). Mas também podemos perceber que todos os grupos sofrem com a presença de outliers.

Figura 8 – Gráfico de boxplot da infraestrutura da quadra do imóvel



Fonte: Elaboração própria

Na tabela 4, é apresentada uma descrição da variável "idade", que representa quanto tempo de construção tem o imóvel, em anos. O valor mínimo do imóvel é 0, indicando que existem imóveis com menos de um ano de construção, já o valor máximo é de 28 anos de idade. A média e a mediana da idade dos imóveis são de 21,98 anos e 25 anos, respectivamente. Percebemos que os valores da média e da mediana são um pouco distantes, indicando que existe outlier nessa variável, com um desvio padrão de 6,86 anos, o que indica uma alta variabilidade nessa variável.

Tabela 4 – Resumo da variável idade

Variável	**Mínimo**	**Mediana**	**Média**	**Máximo**	**Desvio Padrão**
Idade	0	25	21.97876	28	6.85873

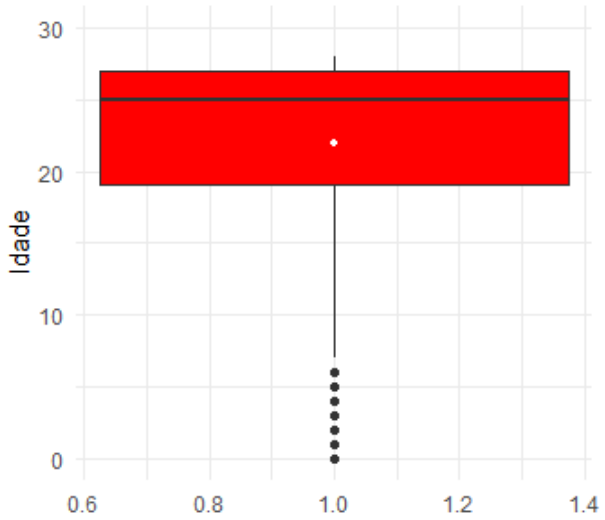
Fonte: Elaboração própria.

Na figura 9, é apresentado o gráfico de boxplot da variável "idade do imóvel". Esse gráfico corrobora com a análise da tabela 4. Podemos perceber, pela diferença interquartil, que essa variável apresenta uma alta variabilidade. Notamos também que existe a presença de outliers baixos, que estão puxando a média para baixo (representada pelo ponto branco).

5.2 Misturas de modelo de regressão linear

Primeiramente, para a construção da mistura de modelos, temos que saber a quantidade de grupos (modelos) que serão gerados. Para isso, foi feita uma análise-teste para encontrar o melhor número de grupos que podem ser gerados. Utilizou-se como base o algoritmo de otimização

Figura 9 – Gráfico de boxplot da idade dos imóveis



Fonte: Elaboração própria

EM (Estimador de Máxima Verossimilhança) para os cálculos, com um número máximo de 10 grupos e 3 repetições.

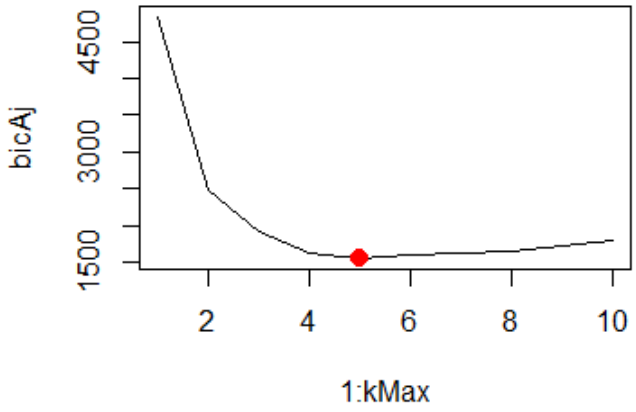
Foi gerada uma mistura de modelos lineares normais do log da variável "valor"como resposta, com todas as variáveis restantes como preditoras. Todos os 10 grupos gerados convergiram; alguns convergiram com poucas interações, enquanto outros, com muitas interações.

Agora, para escolher o melhor grupo de misturas, utilizaremos o critério Bayesiano (BIC) como base para a seleção dos grupos. Podemos observar na Figura 10, que mostra o número do BIC para cada grupo ajustado. Notamos que o menor valor do BIC ocorre quando são cinco grupos. Logo, com base no critério Bayesiano, a melhor divisão de grupos, que são as misturas de modelos gerados, será de cinco.

Na Tabela 5, são apresentados os valores de cada coeficiente de regressão para cada um dos componentes (grupos de mistura de regressão) gerados. Podemos observar que existem 24 coeficientes para cada grupo, sendo que os valores desses coeficientes são distintos entre os componentes, alguns com valores relativamente próximos, como é o caso do coeficiente "ind loc peso0.9", e outros com valores um pouco distantes, como é o caso do coeficiente "stlote".

Fazendo uma análise por componente (grupo), podemos observar que, em alguns coeficientes, a resposta média é negativa apenas para aqueles grupos respectivos, como é o caso do grupo três, onde os coeficientes "idh"e "ind loc peso0.6"têm respostas médias negativas somente nesse componente. Logo, essas variáveis contribuem negativamente para o valor do imóvel nesse grupo, enquanto nos demais grupos, elas contribuem positivamente. Já no grupo cinco, o coeficiente "coordx"tem resposta média positiva somente nesse componente; nos demais,

Figura 10 – Gráfico do valor do BIC por grupo



Fonte: Elaboração própria

ele tem resposta média negativa. Então, essa variável contribui positivamente para o valor do imóvel nesse grupo, enquanto nos demais grupos, ela contribui negativamente.

Já em outros casos, nos grupos, há coeficientes que apresentam valores muito altos, tanto negativos quanto positivos, em comparação com as demais variáveis (coeficientes). Um exemplo é o coeficiente "coordy", onde, no grupo um, quando essa variável aumenta uma unidade, a resposta diminui, em média, 189.98452, enquanto no grupo três, ela aumenta 114.27861. Portanto, podemos perceber que o impacto maior dessa variável, tanto negativamente quanto positivamente, é no grupo um e no grupo três, onde, em um grupo, ela contribui negativamente para o valor do imóvel, enquanto no outro grupo, ela contribui positivamente. O sigma, que representa o desvio padrão no modelo, é relativamente próximo entre os grupos, exceto no grupo três, onde apresentou um sigma igual a 1.15806. Possivelmente, o modelo do grupo três capturou uma heterocedasticidade. O grupo três é o grupo que tem mais coeficientes com impacto negativo, sendo oito (com intercepto) dos 24 coeficientes.

O coeficiente "dist av"tem um efeito muito pequeno em cada um dos grupos, indicando que o impacto dele pode não ser tão influente nos modelos.

Na Tabela 6, é mostrada a probabilidade de um dos grupos e a quantidade de imóveis por grupo. Observamos que, em questão de frequência, o grupo quatro foi o que apresentou mais imóveis, com 3.587 observações, seguido pelo grupo cinco, com 2.173 observações. Enquanto isso, o grupo três foi o que ficou com menos observações, cerca de 122, seguido pelo grupo dois, com 535 observações.

Em relação à análise da probabilidade, que é a probabilidade de uma observação pertencer

Tabela 5 – Coeficientes de Regressão por Componente

Coeficientes	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
(Intercept)	3374.02893	1994.08116	-1778.18060	1910.31781	-569.22687
av_rod1	0.01788	-0.06074	0.08291	-0.00048	0.11785
conservacao3	0.06627	0.09757	0.14491	0.06367	0.04412
coordx	-24.29042	-21.87637	-3.05519	-4.54911	16.29049
coordy	-189.98452	-105.72194	114.27861	-115.01614	22.53729
densidade_vertical	0.09847	0.15207	0.94374	0.20826	0.22380
dist_av	-0.00004	0.00010	-0.00013	0.00019	-0.00020
idade	-0.07445	0.19610	0.14786	0.00265	-0.01653
idh	1.93448	1.83752	-1.42251	0.75188	1.32472
ind_loc_peso0.34	0.12725	0.07122	0.17678	0.03655	0.00219
ind_loc_peso0.6	0.21208	0.16511	-0.04847	0.16592	0.11046
ind_loc_peso0.9	0.22206	0.20911	0.26689	0.27627	0.16356
ind_loc_peso1.15-1.25	0.28086	0.21985	0.10180	0.33560	0.20508
infra_aju2	-0.01348	-0.07677	0.76926	0.06454	-0.04521
infra_aju3	0.40699	-0.11357	-0.05464	0.04236	-0.03681
infra_aju4	0.38054	-0.02632	0.28588	0.09589	0.03554
numpav2-4	0.17154	-0.02420	-0.08829	-0.00296	-0.10041
padrao1-1.2-1.35	0.07524	0.18073	0.21962	0.07620	0.01900
peso2-3	0.02799	0.01951	-0.51520	0.02196	-0.09012
stcunid	0.03111	0.07715	0.00126	0.08026	0.11220
stlote	0.56153	0.28329	0.53642	0.20617	0.09177
tipo_registro_valorOferta	0.52459	0.04283	0.39694	0.26527	0.60577
tipo_registro_valorTransacao	0.08677	0.11398	0.10912	0.07598	0.30106
sigma	0.09324	0.17035	1.15806	0.11002	0.23463

Fonte: Elaboração própria

a um desses grupos, notamos que o grupo cinco foi o que apresentou a maior probabilidade, com quase 35%, seguido pelo grupo quatro, que teve uma probabilidade de pouco mais de 29%. Os grupos que apresentaram menor probabilidade foram os grupos três e dois, com probabilidades de 0,0277% e 0,1546%, respectivamente. Um fator curioso a se observar é que o grupo cinco, que tem a maior probabilidade, tem menos observações que o grupo quatro, que tem uma probabilidade mais de 5% inferior.

Tabela 6 – Probabilidade e frequência por grupo.

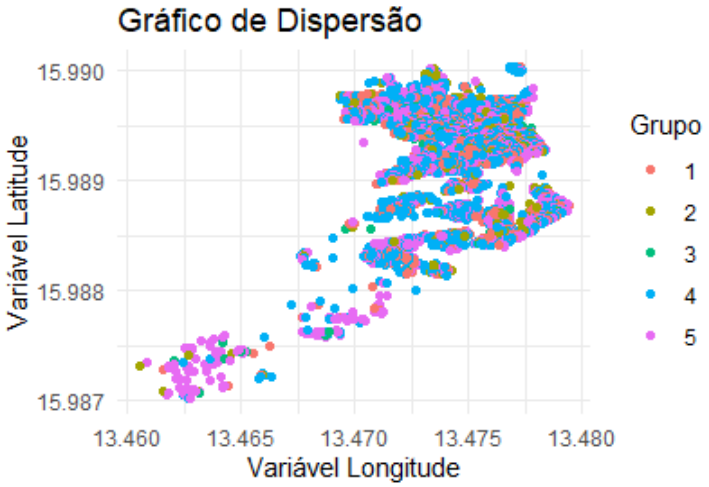
Grupo	**1**	**2**	**3**	**4**	**5**
Probabilidade	0.18455	0.15469	0.02773	0.29156	0.34144
Frequência	1303	535	122	3587	2173

Fonte: Elaboração própria.

Na Figura 11, é apresentado o gráfico de dispersão por grupo. Podemos notar que as cinco cores, que representam cada um dos 5 grupos, estão bem espalhadas por todo o gráfico, indicando que há uma variabilidade alta de cada grupo dentro do município de Aracaju. Isso ocorre porque os eixos X e Y representam a longitude e latitude, e cada ponto representa a

localização do imóvel dentro do município.

Figura 11 – Gráfico de dispersão por grupo



Fonte: Elaboração própria

Na Tabela 7, é apresentada uma descrição do valor do imóvel por grupo. Podemos notar que o grupo três foi o que apresentou o maior desvio padrão, cerca de R$ 5.142.640,00 do valor do imóvel. Isso se deve muito à alta variabilidade dentro dele, onde podemos notar que o valor mínimo foi de R$ 2.651,00 e o máximo foi de R$ 39.500.000,00. O grupo que apresentou menor variabilidade foi o grupo quatro, onde o valor do desvio padrão foi de R$ 209.796,00, com valores de mínimo e máximo de R$ 63.900,00 e R$ 3.300.000,00, respectivamente. O grupo dois foi o que apresentou valores de média e mediana mais próximos, cerca de R$ 303.621,00 para a média e R$ 250.000,00 para a mediana. Em todos os grupos, a mediana foi menor que a média, possivelmente devido à presença de valores de imóveis muito altos (outliers) que acabam puxando essa média para cima.

Tabela 7 – Descritiva do grupo.

Grupo	**Mínimo**	**Mediana**	**Média**	**Máximo**	**Desvio Padrão**
1	39000	270000	382732	19300000	616622
2	37500	250000	303621	2500000	256433
3	2651	362500	1600574	39500000	5142640
4	63900	255000	315107	3300000	209796
5	50000	300000	414110	10000000	439400

Fonte: Elaboração própria.

6 CONCLUSÃO

Por meio deste trabalho, buscou-se construir uma mistura de modelos de regressão linear capaz de precificar o valor dos imóveis em Aracaju-SE, utilizando o algoritmo de misturas de modelos com os dados da SEMFAZ-AJU.

Através da análise descritiva dos dados, conseguimos observar a variabilidade no preço dos imóveis. Identificamos três fontes para a obtenção do valor do imóvel: oferta, transação e ITBI. Percebemos que os preços dos imóveis variavam entre R$2.651 e R$39.500.000, com um desvio padrão de R$764.347. O valor de R$2.651 e outros dois valores semelhantes provavelmente representam erros de digitação, pois é quase impossível um imóvel ter esse preço. O tipo de registro que mais continha observações era o ITBI, com pouco mais de 36% dos dados. Um fator curioso é que, para todos os tipos de registro, o preço médio do imóvel era sempre superior à mediana, o que indica a presença de outliers, que acabam puxando o valor médio para cima.

Quanto à infraestrutura do imóvel, foi analisada a face da quadra em que o imóvel está localizado, considerando os seguintes elementos: água, esgoto, galeria pluvial, guias e sarjetas, iluminação pública e pavimentação. Podemos perceber que 65,48% dos dados apresentam uma soma igual a seis dessas características citadas.

Analisando a variação da variável "ind loc peso", que representa uma aproximação para a localização do imóvel com base na planta de 2014, observamos que a métrica entre 1,15 e 1,25 foi a que apresentou maior variação no valor do imóvel, pois é a que possui o maior intervalo interquartil. Notamos também que, para todos os valores de aproximação da localização do imóvel, existem outliers, tanto altos quanto baixos. Esses são alguns insights que conseguimos perceber no conjunto de dados.

Quanto à aplicação da técnica de misturas de modelos de regressão, conseguimos dividir o conjunto de dados em cinco grupos, gerando cinco modelos de regressão com coeficientes distintos, utilizando o critério Bayesiano como base para selecionar o número de componentes (grupos). A maior probabilidade de uma observação pertencer a um determinado grupo é a do grupo cinco, com uma probabilidade de quase 35%. O grupo com a menor probabilidade é o grupo três, com 0,0277%. Utilizando a latitude e a longitude dos imóveis, podemos notar que todos os grupos estão bem distribuídos no município de Aracaju-SE, sem nenhuma aglomeração de um único grupo em determinado local.

6.1 Trabalhos futuros

Durante a pesquisa, observou-se possíveis erros de mensuração do valor dos imóveis. Seria interessante buscar possíveis correções para o valor desses imóveis, além da inclusão de outros fatores.

A precificação dos imóveis por meio de técnicas de modelagem é bastante utilizada, e a técnica de misturas de modelos de regressão se sobressai entre as demais, como redes neurais, aprendizado de máquina e modelos não paramétricos. Isso ocorre porque, por meio da modelagem de regressão linear, conseguimos medir o impacto de uma determinada variável no modelo, ou seja, como aquela variável impacta no preço do imóvel. Nesse contexto, em trabalhos futuros, surge a necessidade de:

- Análise de Diagnóstico;
- Inferência dos Modelos;
- Interpretação dos Grupos;
- Outras Maneiras de Selecionar o Número de Grupos;
- Avaliar a Capacidade Preditiva do Modelo.

REFERÊNCIAS

ALVES, D. C. de O. et al. Modelagem dos preços de imóveis residenciais paulistanos. **Rev. Bras. Finanças**, v. 9, n. 2, p. 167–187, 2011. Citado na página 17.

CHAGAS, E. T. D. O. Avaliação de imóveis utilizando redes neurais artificiais - revisão sistemática de literatura. **Revista Científica Multidisciplinar Núcleo do Conhecimento**, v. 5, p. 55–75, 2019. Citado na página 17.

FARIA, S.; SOROMENHO, G. Fitting mixtures of linear regressions. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 80, n. 2, p. 201–225, 2010. Citado 3 vezes nas páginas 14, 19 e 20.

GRAYBILL, F. A.; IYER, H. K. **Regression Analysis: Concepts and Applications**. Belmont, CA, EUA: Duxbury Press, 1994. Citado na página 12.

GRUEN, B.; LEISCH, F. **flexmix: Flexible Mixture Modeling**. [S.l.], 2025. R package version 2.3-20. Disponível em: <https://CRAN.R-project.org/package=flexmix>. Citado na página 23.

GRÜN, B.; LEISCH, F. Fitting finite mixtures of generalized linear regressions in r. **Computational Statistics & Data Analysis**, v. 51, n. 11, p. 5247–5252, 2007. Citado na página 13.

HUANG, R. L. M.; WANG, S. Nonparametric mixture of regression models. **Journal of the American Statistical Association**, ASA Website, v. 108, n. 503, p. 929–941, 2013. Citado na página 18.

LEISCH, F. FlexMix: A general framework for finite mixture models and latent class regression in R. **Journal of Statistical Software**, v. 11, n. 8, p. 1–18, 2004. Citado na página 23.

MCLACHLAN, G. J.; KRISHNAN, T. **The EM algorithm and extensions**. 2. ed. Hoboken, NJ, EUA: John Wiley & Sons, 2008. Citado na página 20.

MCLACHLAN, G. J.; PEEL, D. **Finite Mixture Models**. [S.l.]: Wiley, 2004. Citado na página 22.

MEDEIROS, R. de V. V.; CARVALHO, S. T. Modelagem econométrica do preço de alugueis de apartamentos na cidade de petrópolis-rj utilizando regressão linear múltipla. **Revista de Economia da UEG**, v. 13, n. 1, p. 157–174, 2017. Citado na página 17.

NETER, J.; WASSERMAN, W.; KUTNER, M. H. **Applied Linear Regression Models**. Homehood, IL, EUA: R. D. Irwin, 1983. Citado na página 12.

NUNES, D. B.; NETO, J. P. B.; FREITAS, S. M. de. Modelo de regressão linear múltipla para avaliação do valor de mercado de apartamentos residenciais em fortaleza, ce. **Ambiente Construído**, v. 19, n. 2, p. 89–104, 2019. Citado na página 17.

PAIXÃO, L. A. R. Indice de preços hedônicos para imóveis:uma análise para o município de belo horizonte. **Economia Aplicada**, v. 19, n. 1, p. 5–29, 2015. Citado na página 17.

PEREIRA, J. C.; GARSON, S.; ARAÚJO, E. G. Construção de um modelo para o preço de venda de casas residenciais na cidade de sorocaba-sp. **GEPROS. Gestão da Produção, Operações e Sistemas**, v. 7, n. 4, p. 153–167, 2012. Citado na página 17.

PINTO, V. H. L.; FERNANDES, R. A. S. Análise de preços hedônicos no mercado imobiliário residencial de conselheiro lafaiete, mg. **INTERAÇÕES**, Campo Grande, MS, Brasil, v. 20, n. 2, p. 627–643, 2019. Citado na página 17.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>. Citado na página 23.

SANTOS, A. A. S. dos. **Aplicação da teoria de preços hedônicos na análise da influência da distância à praia no valor dos imóveis residenciais em Aracaju/SE**. Monografia (Bacharelado em Engenharia civil) — Universidade Federal de Sergipe, Aracaju, SE, Brasil, 2023. Citado na página 18.

SOUSA, M. M. de. **Modelo Random Forest aplicado a precificação de imóveis à venda em Aracaju, SE**. Monografia (Bacharelado em Estatística) — Universidade Federal de Sergipe, Aracaju, SE, Brasil, 2023. Citado na página 18.