Universidade de Brasília
Departamento de Estatística

Modelagem de dados do mercado imobiliário durante a pandemia do Covid-19 em Brasília

Rafaela M. C. Alonso

Projeto apresentado ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Rafaela M. C. Alonso

Modelagem de dados do mercado imobiliário durante a pandemia do Covid-19 em Brasília

Orientador(a): Prof(a). Leandro Tavares

Projeto apresentado ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Agradecimentos

Aos meus pais, Luana e Marcos e a minha irmã Eduarda, pela confiança no meu progresso e pelo apoio emocional.

Sou grato pela confiança depositada na minha proposta de projeto pelo professor Leandro, orientador do meu trabalho. Obrigado por me manter motivada durante todo o processo.

A todos os meus amigos do curso de graduação que compartilharam dos inúmeros desafios que enfrentamos, sempre com o espírito colaborativo.

Também quero agradecer à Universidade de Brasília e o seu corpo docente que demonstrou estar comprometido com a qualidade e excelência do ensino. Foram anos incríveis de aprendizado que para sempre levarei comigo.

Resumo

Esse trabalho objetiva investigar duas classes de modelos para apresentar o mercado imobiliário da cidade de Brasília, nos bairros Lago Norte e Lago Sul. A primeira classe se trata da elaboração de um modelo utilizando a regressão linear múltipla, a qual permitem estimar os preços dos imóveis considerando as características previamente mapeadas e o grau de correlação entre elas. A segunda visa a elaboração de um modelo utilizando a técnica de modelos lineares generalizados com distribuição gama e função de ligação logarítmica. Este relatório apresenta a aplicação experimental de Ciência de Dados para realizar a avaliação de imóveis a partir de dados coletados em um site de compra e venda. O modelo final apresentou um ajuste adequado aos dados e uma capacidade de previsão bastante satisfatória, tornando-se assim uma ferramenta adicional confiável para avaliação de imóveis.

Palavra-Chave: Modelo linear generalizado; Imóveis; Modelagem; Previsão.

Abstract

This work aims to investigate two approaches to present the real estate market in city of Brasilia, in Lago Norte and Lago Sul neighborhoods. The first approach deals with the elaboration of a model using multiple linear regression, which allows estimating the prices of the properties considering the previously mapped characteristics and the degree of correlation between them. The second approach aims at the elaboration of such a model using the technique of generalized linear models. This report presents the application Experimental Data Science to evaluate real estate from data collected on a buying and selling website. The final model showed an adequate fit data, and a very satisfactory forecasting capacity, thus becoming a reliable additional tool for evaluating urban properties.

Keyword: Generalized linear model; Properties; Modeling; Forecast.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Contexto Pandêmico

A pandemia de Covid-19, causada pelo vírus SARS-CoV-2, tornou-se um dos grandes desafios do século XXI. A Covid-19 é uma doença respiratória causada pelo coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2). A evolução dos casos levou a OMS (Organização Mundial da Saúde) a decretar pandemia em março de 2020.

O Brasil identificou a primeira contaminação pelo novo coronavírus no final de fevereiro de 2020 e, por volta de março do mesmo ano, alguns estados passaram a adotar medidas mais restritivas para o combate da Covid-19, tais como políticas de distanciamento e a restrição de funcionamento de alguns estabelecimentos. Nesse sentido, se inicia uma crise econômica em diversos setores devido a características de queda da oferta e queda da demanda. Contudo, devido ao isolamento, se observou uma nova tendência no mercado imobiliário. Ainda em agosto de 2020, uma nova pesquisa realizada pela Câmara Brasileira da Indústria da Construção (CBIC) em parceria com a Brain Inteligência Estratégica apontou o aumento da intenção de compra de imóveis pelos consumidores, indicando um bom momento para o segmento da construção, o qual retornou a níveis de antes da pandemia (CBIC, 2020).

A busca por uma nova residência era pautada por características não tão valorizadas antes. Com o advento do trabalho à distância, e a não obrigatoriedade do deslocamento a sede das empresas, os compradores passaram a dar maior preferência a moradias com maior área aberta, em detrimento de uma residências perto do local de trabalho. Contudo, os seguintes fatores corroboraram para que o segmento pudesse se manter firme e com boas perspectivas:

- Queda na taxa de juros;
- Bom preço para compradores e vendedores;
- Planejamento de longa data e crédito mais acessível.

Enquanto o Banco Central esperava uma queda de 6,5% para a economia brasileira, o setor imobiliário iniciou um progresso com a redução histórica da Selic para 2%. Este menor patamar da taxa básica de juros criou condições favoráveis para as operações de crédito e investimentos, o que refletiu nas vendas do setor. Dados do Boletim de Conjuntura do Sindicato da Habitação do Distrito Federal (Secovi/DF), por sua vez, mostram evolução de 40,5% no valor de financiamentos imobiliários nos sete primeiros meses de 2020 em relação ao mesmo período do ano de 2019 (JOVEMPAN, 2020).

Com o objetivo de estimular o setor, a Caixa Econômica federal anunciou a substituição da pausa no pagamento do financiamento imobiliário, concedido no início da pandemia, pela possibilidade de o comprador pagar apenas 50% da mensalidade por até 3 meses, ou de 50% a 75% por até 6 meses.

1.2 Mercado Imobiliário

O mercado imobiliário é considerado um setor onde ocorrem as transações de imóveis. Segundo as definições de mercado apresentadas por Matos e Bartkiw (2013), tal setor pode ser caracterizado como um conjunto de compradores e vendedores que atuam interagindo com a finalidade de comprar ou vender seus produtos ou serviços. Nele, atuam pessoas físicas ou jurídicas, e também as imobiliárias. Sua atuação pode ocorrer por meio de um aluguel, compra ou venda de um bem imóvel. O Código Civil diz que:

> São bens imóveis o solo e tudo quanto se lhe incorporar natural ou artificialmente.(BRASIL,2022,Art 79).

De acordo com Arraes e Filho (2008), o consumo de habitação é inerente a todo ser humano, sendo caracterizado como necessidade básica e intimamente ligado à busca de segurança contra as adversidades do meio ambiente. Porém, o consumo de habitação pode ser segmentado em dois grandes grupamentos: aqueles que possuem a intenção de utilizar o bem para satisfação final de sua necessidade básica de habitação e aqueles que o adquirem para compor cesta de bens de investimento.

Ademais, Brito e Rodrigues (2014) relata a aquisição da casa própria como um papel relevante em nossa sociedade. Tal ato está fortemente ligado a aspectos culturais os quais legitimam essa dívida como prioritária e fundamental. Para ele, a razão da função da propriedade está fortemente presente no imaginário coletivo brasileiro.

A aquisição de um imóvel, por meio de financiamento, assume status de comprometimento financeiro legitimado de um imóvel. Assim, por meio da expansão do crédito, houve uma maior acessibilidade para a aquisição de financiamentos habitacionais das famílias brasileiras. Nesse sentido, faz-se presente estudos que buscam estimar os valores dos imóveis tendo em vista variáveis como: área do terreno, número de quartos e localização. Além disso, para a realização da análises, deve-se considerar que o valor do imóvel (Y), o qual representa a variável resposta, deve ser previsto, e que o conjunto de potenciais variáveis explicativas inclui todas as demais variáveis do conjunto de dados, com exceção da variável de identificação.

Assim, o mercado imobiliário, pela sua complexidade, impõe a necessidade de resultados precisos e confiáveis na quantificação dos valores imobiliários. Para alcançar

estes objetivos é necessária a adoção de procedimentos fundamentados, que sejam baseados em métodos científicos, que minimizem a parcela de subjetividade existente na formação do valor. Nesse sentido, o presente trabalho apresenta uma aplicação estatística no mercado imobiliário, cujo objetivo é propor um modelo estatístico para previsão de preço de imóveis e quantificar a influências de suas características nessa precificação. Para tal estudo, será utilizado um banco de dados de 2020 e 2021 para as regiões do Lago Norte e Lago Sul, ambas setorizadas em Brasília.

Como objetivos específicos, pode-se destarcar:

- Verificar a relação entre as vendas do imóvel e da localização por meio de uso de modelos de regressão linear múltipla e lineares generalizados;
- Comparar os modelos com relação ao tipo de regressão empregada;
- Analisar a influência do decorrer da pandemia em relação aos elementos que influenciam no modelo, por meio da inserção da variável "ano"no modelo;
- Desenvolver uma plataforma interativa para as variáveis utilizadas;

2 Referencial Teórico

2.1 Regressão Linear Múltipla

Os modelos de regressão estruturam-se em uma coleção de técnicas estatísticas com a função de descrever a relação entre uma variável de interesse com uma ou mais variáveis explicativas. Quando tal modelo contém mais de um regressor é chamado de modelo de regressão linear múltipla.

O modelo o qual associa a variável dependente com as variáveis independentes pode ser expresso da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_p + \epsilon, \tag{2.1.1}$$

onde $\beta_i, i = 0, 1, ..., \mathbf{k}$ são os parâmetros a serem estimados com $\mathbf{k}$ variáveis regressoras, ϵ é uma variável aleatória desconhecida, conhecida como erro aleatório, a qual interfere no resultado das observações da variável dependente Y, tal que $\epsilon \sim N(0, \sigma^2)$ e $x_i, i = 0, 1, ..., \mathbf{p}$ é nossa variável preditora. O termo linear é usado porque a equação 2.1.1 é uma função linear dos parâmetros desconhecidos $\beta_0, \beta_1, ..., \beta_{\mathbf{k}}$.

Temos então o seguinte sistema escrito em notação matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{11} & \cdots & x_{11} \\ 1 & x_{12} & x_{11} & \cdots & x_{11} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix} \Leftrightarrow y = X\beta + \varepsilon. \tag{2.1.2}$$

Com a utilização do banco de dados estimamos $\beta_0, \beta_1, ..., \beta_k$, iremos substituir estes parâmetros pelas suas estimativas $b_0, b_1, ..., b_k$ para obter a equação de regressão estimada. Nesse sentido, o método de mínimos quadrados, representado na Figura 1, pode ser utilizado com o intuito de estimar o modelo. O ajuste dos parâmetros por tal método consiste em determinar os valores que minimizam a soma dos quadrados das diferenças entre o valor estimado e os dados experimentais (HELENE, 2006). Assim, devemos minimizar:

$$\sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 x_{1i} - ... - \beta_k x_{ki})^2. \tag{2.1.3}$$

O mínimo de $\sum_{i=1}^{n} (\varepsilon_i)^2$, também conhecido como resíduo, é obtido derivando-a

em relação a $\beta_0, \cdots, \beta_k$ e igualando o resultado a zero. Sendo assim, para cada observação $(y_i, , x_{1i}, , \cdots, x_{ki})$ temos o seguinte resíduo:

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1i} + ... + b_k x_{ki}). \tag{2.1.4}$$

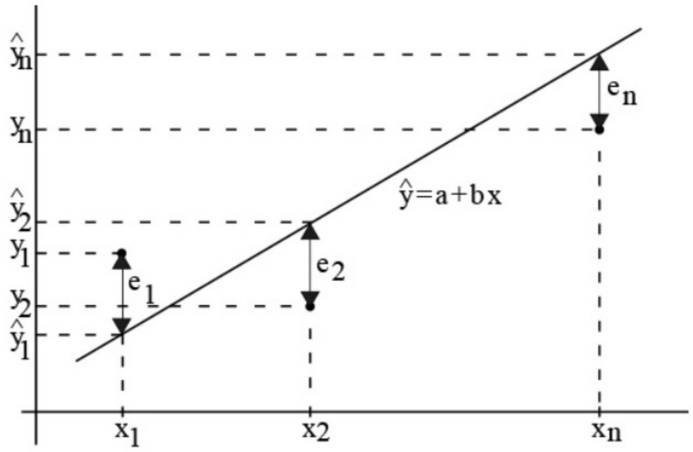


Figura 1: Representação dos desvios para o método de Mínimos Quadrados
Fonte: Martins, E.G.M., (2019) Regressão linear simples, Rev. Ciência Elem., V7(3):04

Uma medida útil associada ao modelo de regressão é o grau em que as predições baseadas na equação $\hat{y}_i$ superam as predições baseadas em y. Se a dispersão (erro) associada a equação é significativamente menor que a dispersão associada a $\bar{y}$, as predições baseadas no modelo serão melhores que as baseadas em $\bar{y}$.(HENRIQUES, 2011). Assim, pelo Método de Mínimos quadrados tem-se que:

$$\hat{\beta} = (X'X)^{-1} Xy, \tag{2.1.5}$$

desde que $(X'X)^{-1}$ exista.

A soma dos quadrados do total de y (SQT) pode ser relacionada com a soma de 2 variáveis: soma dos quadrados da regressão (SQReg), e a soma dos quadrados dos resíduos (SQE). Assim, tem-se:

$$SQT = SQReg + SQE. \tag{2.1.6}$$

Trocando para os valores correspondentes para cada uma das somas, tem-se:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (\hat{y}_i - y_i)^2. \tag{2.1.7}$$

Logo, para o caso de uma regressão linear múltipla, temos para o Quadrado Médio

da Regressão (QMR_{eg}) e Quadrado Médio do Resíduo (QMR_{es}):

$$QMR_{eg} = \frac{SQR_{eg}}{k-1},$$

e

$$QMR_{es} = \frac{SQE}{n-k}.$$

A Tabela ANOVA fica da seguinte forma :

Tabela 1: Tabela ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio
Regressão	$k-1$	$\hat{\beta}'X'Y - n\bar{Y}^2$	QMR_{eg}
Resíduo	$n-k$	$YY' - \hat{\beta}'X'Y$	QMR_{es}
Total	$n-1$	$YY' - n\bar{Y}^2$	

O quociente entre o SQReg e SQT nos da uma medida da proporção da variação total a qual é explicada pelo modelo de regressão, sendo essa medida conhecida como coeficiente de determinação. Tal quociente é representado por:

$$r^2 = \frac{SQReg}{SQT} = \frac{SQT - SQE}{SQT} = \frac{SQT}{SQT} - \frac{SQE}{SQT} = 1 - \frac{SQE}{SQT}. \tag{2.1.8}$$

Temos então $0 \leq r^2 \leq 1$. Sendo assim, temos as seguintes interpretações:

- $r^2 \cong 0$: significa que grande parte da variação de Y não é explicada linearmente pelas variáveis independentes;
- $r^2 \cong 1$: significa que parte da variação de Y é explicada linearmente pelas variáveis independentes.

A raiz quadrada de r^2, para o caso em que estão envolvidas pelo menos duas variáveis independentes, é chamada de coeficiente de correlação múltiplo. Ela é uma medida do grau de associação linear entre Y e o conjunto de variáveis $X_1, X_2, ..., X_k$. Desse modo, temos as seguintes interpretações:

- $r = 0$: indica a inexistência de qualquer relação linear entre a variável dependente Y e o conjunto de variáveis independentes $X_1, X_2, ..., X_k$;
- $r \neq 0$: indica a existência de uma associação linear entre a variável dependente Y e o conjunto de variáveis independentes $X_1, X_2, ..., X_k$. Assim, Y pode ser expresso como uma combinacão linear de $X_1, X_2, ..., X_k$.

Nesse sentido, com o intuito de verificar a adequação do modelo de regressão, é utilizada a análise de resíduos.

Para tal os resíduos devem verificar os pressupostos:

- $\varepsilon_i, i = 1, ..., n$ são normalmente distribuídos;
- $var(\varepsilon_i) = \sigma^2, i = 1, ..., n$ tem variância constante (homoscedasticidade);
- ε_i e ε_j são independentes, para todo $i \neq j$.

Para a verificação das suposições acima, sertão utilizadas as seguintes práticas:

- Teste de Kolmogorov-Smirnov;
- Análise do gráfico dos resíduos versus valores ajustados;
- Teste de Durbin-Watson.

2.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados(MLGs) são uma extensão do modelo linear

$$Y = X\beta + \epsilon, \tag{2.2.1}$$

quando a suposição de normalidade não é plausível para o erro do modelo. Tal modelo possibilita utilizar outras distribuições para os erros e uma função de ligação relacionando a média da variável resposta à combinação linear das variáveis explicativas(OLIVEIRA, 2019).

Cordeiro e Demétrio (2008) sugerem que os MLG são caracterizados pela seguinte estrutura:

1. O componente aleatório de um MLG é definido a partir da família exponencial uniparamétrica, ou multiparamétrica, na forma canônica com a introdução de um parâmetro $\phi > 0$ de perturbação, que é uma medida de dispersão da distribuição. Assim, é mencionado que:

$$f(y; \theta, \phi) = \exp\left\{\phi^{-1}[y\theta - b(\theta)] + c(y, \phi)\right\}, \tag{2.2.2}$$

   em que b(·) e c(·) são funções conhecidas.

2. A variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família de distribuições 2.2.2 que engloba a distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens;

3. A ligação entre os componentes aleatórios e sistemáticos é feita por meio de uma função adequada denominada função de ligação.

Relacionando modelo linear com MLG, pode-se entender que a função de ligação possui um papel similar a uma transformação na resposta do modelo linear de regressão (ASEVEDO, 2011). Uma função de ligação transforma as probabilidades dos níveis de uma variável de resposta categórica em uma escala contínua que é ilimitada. Depois de concluída a transformação, a relação entre os preditores e a resposta pode ser modelada com regressão linear. Sendo assim, a forma geral de tal função é:

$$g(\mu_i) = X_i\beta, \tag{2.2.3}$$

sendo β o vetor dos coeficientes de regressão associados ao preditor.

Nos MLG's utiliza-se a maximização da função de log- verossimilhança, descrita na equação 2.2.4, para se obter as estimativas dos coeficientes. Assim, a função de log-verossimilhança é representada por:

$$l = ln\, l(y;\theta,\phi) = \sum_{i=1}^{n} ln\, f(y_i;\theta_i,\phi) = \sum_{i=1}^{n} \left\{ \frac{[y_i\theta_i - b(\theta_i)]}{\phi_i} + c(y_i,\phi) \right\}. \tag{2.2.4}$$

Porém, a maximização dessa função depende da utilização de métodos numéricos. Para resolver este problema, é indicado o processo de otimização iterativo chamado de algoritmo de Newton-Raphson, definido como:

$$\hat{\beta}^{(m+1)} = (X'W^{(m)}X)^{-1}X'W^{(m)}z^{(m)}, \tag{2.2.5}$$

onde $\hat{\beta}^{(m+1)}$ é a estimativa do vetor de parâmetros na iteração atual; X é, como de costume, a matriz do modelo, dos valores das variáveis de regressão; W é a matriz de pesos diagonal e z é o vetor das variáveis de ajuste na m-ésima iteração. Assim, de acordo com Asevedo (2011), pode-se dizer que o algoritmo inicia o processo especificando uma estimativa inicial e vai sucessivamente alterando-a até que a diferença entre o β na iteração $(m+1)$ e a estimativa anterior seja menor que um γ pré-definido, sendo obtida assim a convergência na matriz dos coeficientes estimados.

2.3 Variáveis Dummy

A primeira etapa na formulação de uma equação de regressão com parâmetros variáveis consubstancia-se na tomada de posição acerca da forma como se processam essas variações. Tem-se a equação correspondente:

$$y_i = \beta_0 + \beta X_i + \epsilon_i, \quad i = 1, 2, ..., n, \quad \epsilon_i \sim N(0, \sigma^2). \tag{2.3.1}$$

Neste contexto, admita-se que da estimação do modelo de regressão, com base numa amostra de n observações, resultava um valor significativo para a estatística t associada ao coeficiente da variável X e, simultaneamente, um valor para o coeficiente de determinação (r^2) relativamente baixo. Estes resultados fazem pensar que embora as variáveis preditoras constituam de facto um fator determinante no comportamento de y, existe ainda uma forte componente não explicada na variabilidade de y, ou dito de outra forma, que o modelo pode encontrar-se mal especificado por incorreta omissão de variáveis explicativas (MISSIO; JACOBI, 2007). Nesse sentido, segue a figura representativa abaixo a qual retrata a relação entre a variável dependente e o regressor X.

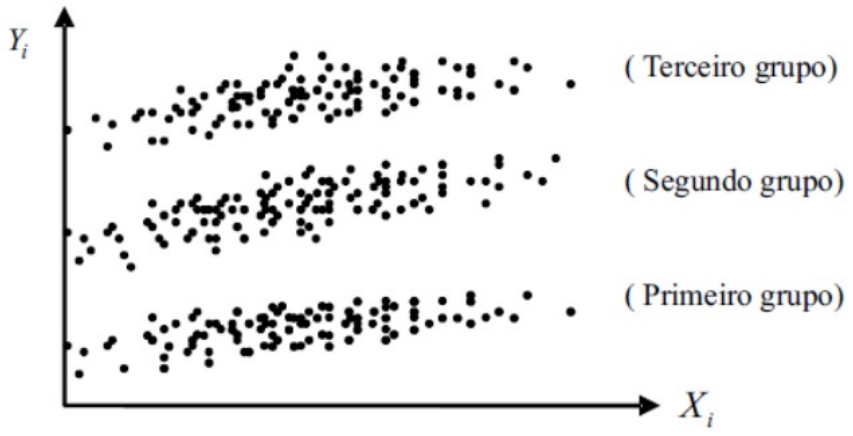


Figura 2: Gráfico de Dispersão entre a variável dependente e o regressor X
Fonte: Valle e Rabelo (2002)

A observação da Figura 2 demonstra um relacionamento positivo entre as variáveis em análise. Contudo, parece existir uma relação distinta entre as variáveis para as observações que pertençam a grupos distintos. Sendo assim, uma forma de solucionar tal problemática seria considerar, em separado, cada um dos grupos de observações e utilizá-los para ajustar três modelos de regressão distintos. De acordo com Missio e Jacobi (2007), as retas de regressão que melhor se ajustam às nuvens de pontos parecem diferir apenas no termo independente pelo que, em termos formais, a sua estrutura deverá ser:

- $y_1 = \beta_{0_1} + \beta X_i + \epsilon_i$ para o primeiro grupo;

- $y_2 = \beta_{0_2} + \beta X_i + \epsilon_i$ para o segundo grupo;
- $y_3 = \beta_{0_3} + \beta X_i + \epsilon_i$ para o terceiro grupo.

Contudo, da estimação dos modelos acima não resultará necessariamente em um mesmo valor para β. A definição de regressores dummy apresenta-se como o procedimento adequado à prossecução deste objectivo.

Com efeito, a definição das variáveis:

$$D_{2i} = \begin{cases} 1, \text{ se a observação verifica a característica que define o segundo grupo;} \\ 0, \text{ caso contrário.} \end{cases} \tag{2.3.2}$$

$$D_{3i} = \begin{cases} 1, \text{ se a observação verifica a característica que define o segundo grupo;} \\ 0, \text{ caso contrário.} \end{cases} \tag{2.3.3}$$

Note que:

Se $D_{2i} = D_{3i} = 0$, o modelo reduz-se a:

$$y_i = \beta_1 + \beta X_i + \mu_i. \tag{2.3.4}$$

Se $D_{2i} = 1$ e $D_{3i} = 0$,

$$y_i = \beta_1 + \beta X_i + \mu_i. \tag{2.3.5}$$

Se $D_{2i} = 0$ e $D_{3i} = 1$,

$$y_i = \beta_3 + \beta X_i + \mu_i. \tag{2.3.6}$$

Daí, ajusta-se o novo modelo de regressão,

$$y_i = \beta_{0_1} + (\beta_{0_2} - \beta_{0_1})\, D_{2i} + (\beta_{0_3} - \beta_{0_1})\, D_{3i} + \beta_1 X_i + \epsilon_i, \tag{2.3.7}$$

em que $i = 1, 2, ..., n$, $\epsilon_i \sim N(0, \sigma^2)$, a mesma estimativa para β e interceptos diferentes. A situação oposta também pode ocorrer. As retas de regressão podem ter o mesmo intercepto com coeficientes de inclinação distintos.

2.4 Análise de Cluster

A análise de cluster objetiva solucionar o seguinte problema: dada uma amostra de n objetos, cada um dos quais caracterizados por p variáveis, devemos criar um critério para se agrupar os objetos em classes, de forma que objetos que possuam características semelhantes estejam na mesma classe. O método deve ser quantitativo e o número de classes não é conhecido.

O método de clusterização K-means classifica os objetos dentro de múltiplos grupos, de forma que a variação intra-cluster seja minimizada pela soma dos quadrados das distâncias Euclidianas entre os itens e seus centroides, o qual é representado por:

$$W\left(C_m\right) = \sum_{x_{i \in C_m}} \left(x_{i-\mu_m}\right)^2 . \tag{2.4.1}$$

Desta forma x_i é o ponto que pertente ao cluster C_m e μ_m representa a média do valor atribuído ao cluster C_m. Cada observação (x_i) é designada a um cluster de forma que a soma dos quadrados da distância da observação em relação ao seu cluster central (μ_m) é mínima 2.4.2 (KASSAMBARA, 2017). Ainda, para definir a variação intra-cluster é utilizada a fórmula a seguir:

$$tot.intracluster = \sum_{m=1}^{m} W\left(C_m\right) = \sum_{m=1}^{m} \sum_{x_i \in C_m} \left(x_{i-\mu_m}\right)^2 . \tag{2.4.2}$$

2.5 Critério de Informação

No processo de mensurar a qualidade de um modelo estatístico, diferentes critérios podem ser usados para comparar os modelos produzidos. Busca-se pelo modelo mais parcimonioso, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável resposta. O conceito de melhor modelo é controverso, mas um bom modelo deve conseguir equilibrar a qualidade do ajuste e a complexidade, sendo esta, em geral, medida pelo número de parâmetros presentes no modelo; quanto mais parâmetros, mais complexo o modelo, sendo pois mais difícil interpretar o modelo. Dentre os critérios para seleção de modelos, os critérios baseados no máximo da função de verossimilhança (MFV) são os mais utilizados, com maior ênfase o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC).

O critério de informação de Akaike (Akaike Information Criterion), relata que o viés é dado assintoticamente por h, em que h é o número de parâmetros a serem estimados

no modelo. Ele é definido por:

$$AIC = -2log\, l(y; \theta, \phi) + 2h, \tag{2.5.1}$$

em que $\log(\hat{\theta})$ é a log-verossimilhança maximizada do modelo. Menores valores de AIC representam uma maior qualidade e simplicidade.

Outrossim, o Critério Bayesiano de Schwarz, proposto por Schwarz (1978), é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O valor é dado por:

$$BIC = -2 \log l(x_h|\theta) + k \log n, \tag{2.5.2}$$

sendo $l(x_n|\theta)$ o modelo, h o número de parâmetros a serem estimados e n o número de observações da amostra. O modelo com menor BIC é considerado o de melhor ajuste.

Contudo, a estimativa do erro de previsão é necessária para avaliar a desempenho dos modelos montados. A separação dos dados em duas partes disjuntas pode trazer resultados divergentes, dependendo da informação contida em cada conjunto. Assim, a validação cruzada é amplamente usado para estimar o erro de previsão.

2.6 Validação Cruzada

A abordagem de validação cruzada por k-fold consiste em dividir os dados de entrada em w partes iguais, também conhecidos como folds. Ajusta-se o modelo utilizando $w - 1$ partes, e a parcela restante fica destinada à validação. Esse processo é repetido w vezes com um subconjunto diferente reservado para avaliação. Por fim, os resultados são combinados obtendo a médias dos erros obtidos.

Figura 3: Validação cruzada k-fold

Assim, sejam K partes denotadas por $C_1, C_2, \ldots, C_w$, em que C_w representa o índice da w-ésima parte e, considerando a presença de $n \cdot w$ observações na partição w, temos que a estimativa do erro de teste é representada por:

$$C_{(W)} = \sum_{k=1}^{K} \frac{nw}{n} EQM_w, \tag{2.6.1}$$

em que

$$EQM_w = \frac{\sum_{i \in C_w} (y_i - \hat{y}_i)^2}{n_w}, \tag{2.6.2}$$

e $\hat{y}_i$ é o valor ajustado da observação i, obtido dos dados com a w-ésima parte removida.

3 Materiais e método

3.1 Conjunto de dados

Os dados utilizados no presente estudo foram fornecidos pela imobiliária brasiliense **61 imóveis** a qual possui um site de vendas de imóveis. Para a análise em questão, será utilizado um conjunto de dados com 23356 imóveis coletados durante o ano de 2020 e 2021, os quais passarão por uma filtragem e seleção. A filtragem consiste em selecionar casas localizadas nas regiões administrativas do Lago Norte e Lago Sul do Distrito Federal, bem como retirar domicílios os quais contenham informações ou valores não condizentes. Após a filtragem, o banco de dados final consiste em 8867 imóveis. Assim, foram retirados os dados que continham as seguintes informações:

- Preço de venda inferior a 300 mil ou superior a 30 milhões;
- Valor do metro quadrado inferior a 20 ou superior a 30 mil reais;
- Área inferior a $100m^2$ ou superior a 6000 m^2;
- Número de suítes superior a 8;
- Número de vagas superior a 8;
- Número de quartos superior a 8;
- Número de suítes superior ao número de quartos;

Na tabela 2 é apresentada as variáveis de estudo para a realização do modelo:

Tabela 2: Banco de dados

Nº	Variável	Nome	Descrição
1	Y	preço	Valor de venda do imóvel (em reais)
2	x1	área	Área(m^2) do imóvel
3	x2	quartos	Número de quartos
4	x3	suítes	Número de suítes
5	x4	vagas	Número de vagas de garagem
6	x5	ano	Ano de inserção do imóvel na plataforma
7	x6	mês	Mês de inserção do imóvel na plataforma
8	x7	bairro	Bairro do local do imóvel
9	x8	quadra	Quadra do local do imóvel

Supondo que a distribuição da variável em questão possui um viés, ou seja, uma das extremidades elevadas e uma cauda longa, medidas como correlação ou regressão podem ser bastante influenciadas pelo pico da distribuição, outliers, dentre outros. Assim,

com o intuito de auxiliar a criação dos modelos, a aplicação da transformação logaritma será aplicada na variável **preço**. Por fim, para o modelo MLG a função de ligação logarítmica será utilizada para relacionar a variável resposta às variáveis explicativas. Essa função é muito importante para essas regressões, pois impede o surgimento de resultados negativos e fornece boas interpretações a partir do exponencial dos coeficientes.

3.2 Sistema Computacional

O SAS Visual Data Mining and Machine Learning (VDMML) é uma solução da Plaforma SAS Viya representado pelo componente ou ação "Build Model" do SAS Drive. O SAS Model Studio é a interface web central que contém o SAS VDMML e outras soluções. O SAS VDMML possui uma interface web projetada para produzir as etapas analíticas de ponta a ponta, tais como: processo de preparação dos dados/manipulação análise exploratória, construção de novas variáveis (engenharia de recurso), técnicas modernas de estatística, mineração de dados e aprendizado de máquina em uma ambiente de processamento in-memory . O Model Studio foi projetado para aproveitar os ambientes de programação e processamento em nuvem do SAS Viya. Nesse sentido, tal sistema foi utilizada com o objetivo de criar os modelos, bem como de criar uma plataforma interativa com os resultados finais.

4 Resultados

4.1 Análise Exploratória

A análise exploratória dos dados tem como objetivos avaliar descritivamente a correlação entre as medidas repetidas, a natureza da tendência temporal, a heterogeneidade dos dados e a presença de valores extremos (outliers). As ferramentas utilizadas na análise exploratória são técnicas gráficas e descritivas.

A correlação entre as variáveis com a componente do valor de venda do imóvel será medida por meio do coeficiente de correlação de Pearson, o qual varia entre -1 e +1, cujos valores próximos de -1 e +1 indicam forte correlação linear e próximos de 0 indicam ausência de correlação linear. Tal relação esta representada na tabela a seguir:

Tabela 3: Relação das variáveis de estudo com preço

Variáveis	Correlação
Área	0,33
Quartos	0,24
Suítes	0,44
Vagas	0,17
Ano	0,07

4.1.1 Preço

A variável **preço** foi filtrada para conter os dados com imóveis que são avaliados entre 300 mil e 30 milhões de reais, possuindo como média 4,1 milhões. Assim, aplicando no dado de estudo, tal coluna possui a maioria dos casos (80%) entre 1,8 milhões e 7,5 milhões e com a variável preço diferenciando melhor nos casos mais altos (10% superiores) e mais baixos (10% inferiores). Por fim, há 619 casos que podem ser valores atípicos, com valor maior que ou igual a 9,2 milhões. Contudo, tais valores ao serem analisados em conjunto com as outras variáveis explicativas, não são considerados outliers.

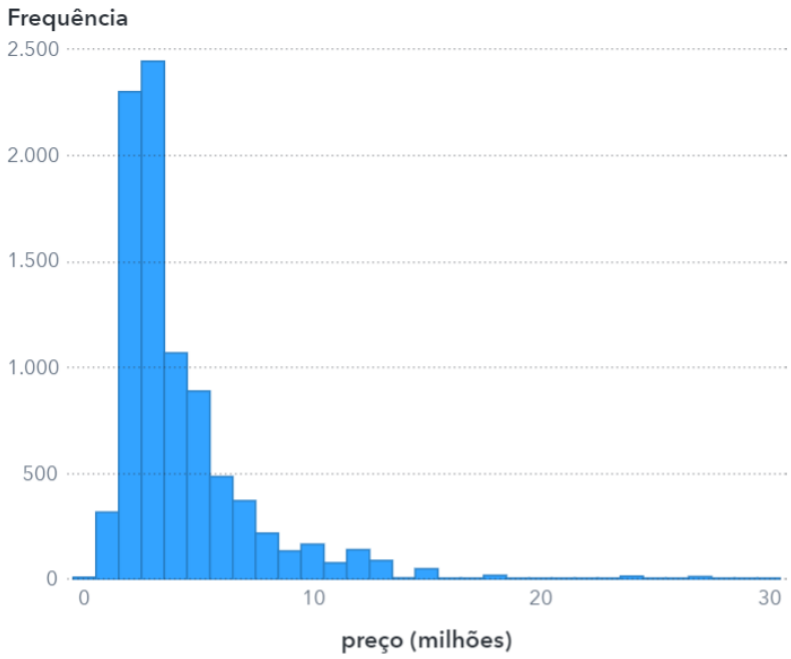


Figura 4: Distribuição de frequência do preço (em Reais)

No caso da variável preço na escala logarítmica, se obtém o seguinte resultado:

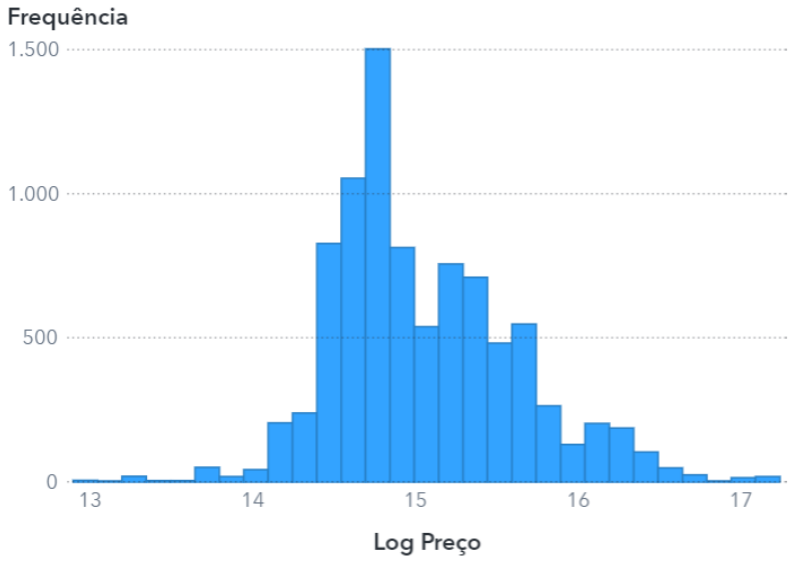


Figura 5: Distribuição de frequência da variável $log\,(preco)$

Observando a Figura 5, é notório o fato de que, utilizando a transformação logarítmica para a variável resposta, os dados passaram a seguir uma distribuição mais simétrica.

4.1.2 Área

A variável **área** foi filtrada para conter os dados com imóveis que possuem entre $100m^2$ e $6000m^2$, possuindo como média $716m^2$. Assim, aplicando no dado de estudo, tal coluna possui a maioria dos casos (80%) entre $384m^2$ e $1100m^2$ e com a variável preço diferenciando melhor nos casos mais altos (10% superiores) e mais baixos (10% inferiores). Por fim, há 369 casos que podem ser valores atípicos, com área maior que ou igual a $1400m^2$. Contudo, tais valores ao serem analisados em conjunto com as outras variáveis explicativas, não são considerados outliers.

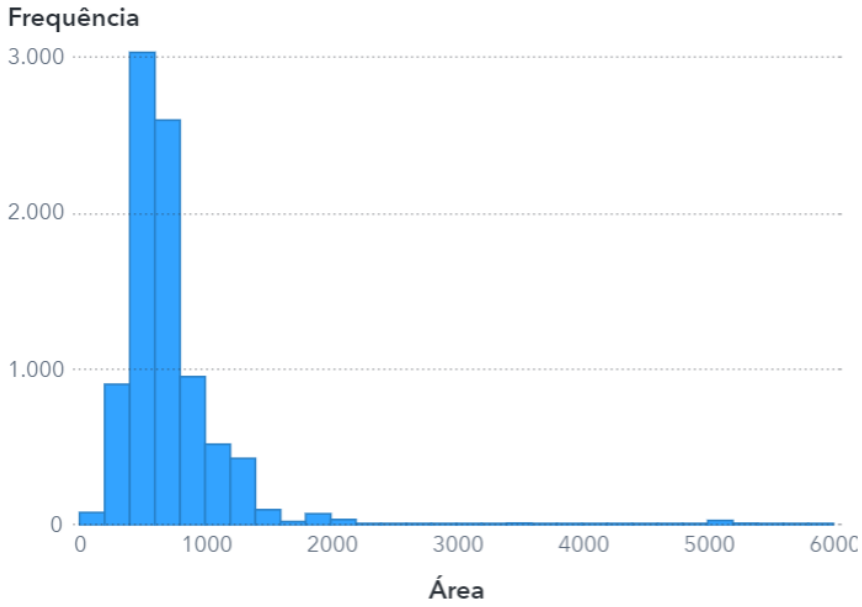


Figura 6: Distribuição de frequência da variável Área

O resultado observado para o coeficiente de correlação de Pearson em relação a variável **preço** é de 0,33, indicando uma relação positiva a qual pode ser observada no gráfico a seguir:

Observando a Figura 7, constata-se que, apesar de uma relação positiva entre as variáveis, os imóveis com o preço mais elevado não são necessariamente os que possuem uma maior área, indicando assim o fato que, outras variáveis também influenciam no preço. Além disso, há a presença de residências com uma elevada área porém com o preço não tão elevado, reafirmando assim a questão de essa não ser a única variável explicativa.

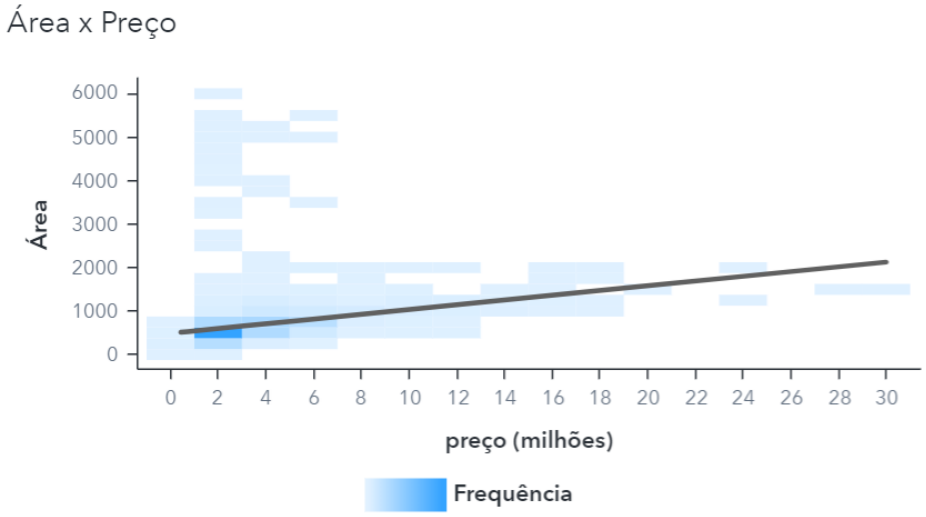


Figura 7: Gráfico de Calor de área por preço

4.1.3 Quartos

A variável **quartos** foi filtrada para conter os dados com imóveis que possuem de 1 até 8 quartos. Tal informação possui a maioria dos casos(81%) entre 4 e 6 quartos e com uma média de 4,6. Por fim, há 429 casos que podem ser outliers. Desses, 388 se encontram com um valor igual ou maior que 7 e, 41 possuem 2 quartos ou menos.

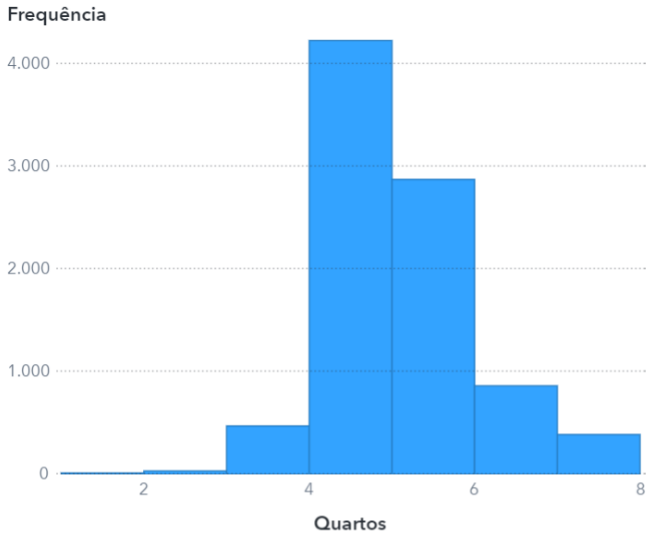


Figura 8: Distribuição de frequência da variável Quartos

O resultado observado para o coeficiente de correlação de Pearson em relação a variável **preço** é de 0,24, indicando uma relação positiva.

4.1.4 Suítes

A variável **suítes** foi filtrada para conter os dados com imóveis que possuem entre 0 e 8 suítes, além de, o número de suítes deve ser menor ou igual ao número de quartos. Tal informação possui a maioria dos casos(80%) entre 2 e 5 suítes. Os três valores mais relacionados são: quarto, valor e área.

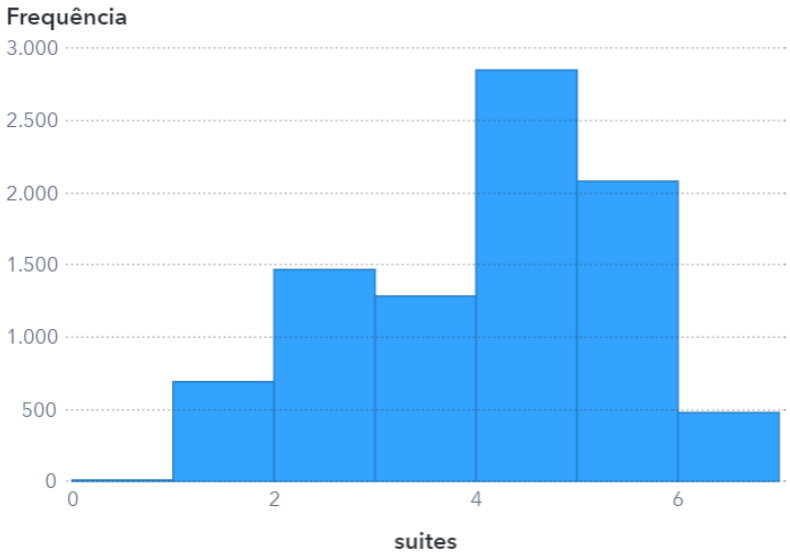


Figura 9: Distribuição de frequência da variável Suítes

O resultado para o coeficiente de correlação de Pearson em relação a **preço** é de 0,44, indicando uma relação positiva a qual pode ser observada no gráfico a seguir:

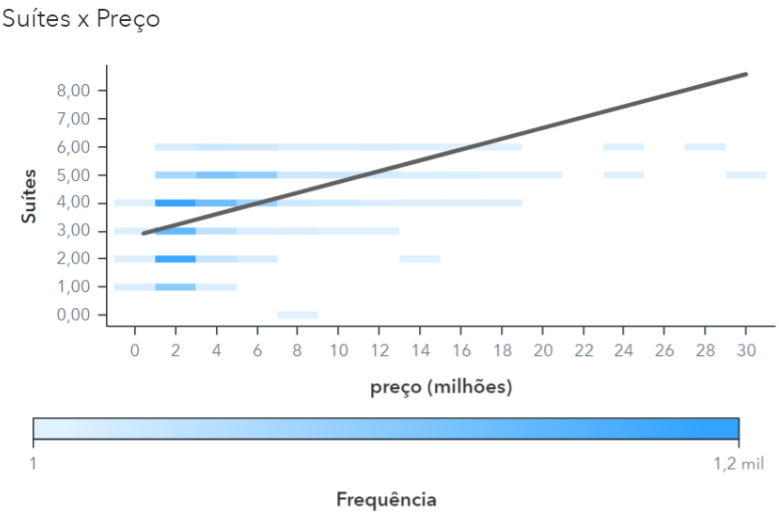


Figura 10: Gráfico de Calor de suíte por preço

4.1.5 Vagas

A variável **vagas** foi filtrada para conter os dados com imóveis que possuem entre 1 e 8 vagas. Tal informação possui a maioria dos casos(79%) entre 2 e 6 vagas. Os fatores mais relacionados são: quadra, valor e suítes.

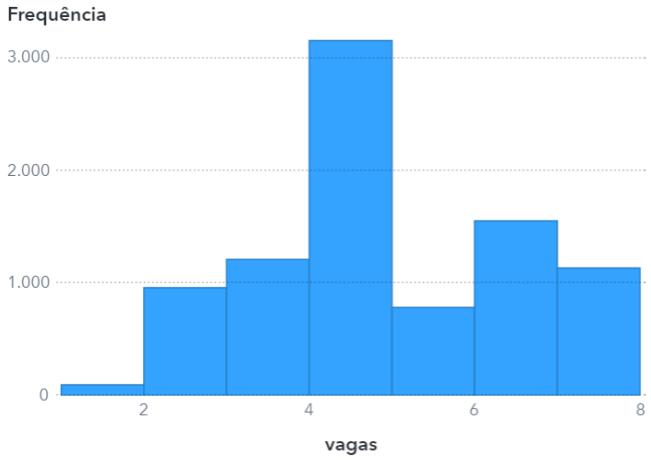


Figura 11: Distribuição de frequência da variável Vaga

O resultado observado para o coeficiente de correlação de Pearson em relação a variável **preço** é de 0,17, indicando uma relação positiva fraca.

4.1.6 Bairro

Por fim, para a variável a qual relata a quadra de localização do imóvel, se foi utilizado os pontos de latitude e longitude a fim de realizar uma visualização por meio de um mapa.

Figura 12: Mapa relacionando a variável preço por localização

4.2 Tratamento dos Dados

Ao analisar uma base de dados, um dos principais desafios do analista é resumir a informação coletada. Em muitos casos, quando contamos com um grande número de observações, pode ser de interesse criar grupos. Dentro de cada grupo os elementos devem ser semelhantes entre si e diferentes dos elementos dentro dos outros grupo.

Nesse sentido, a tabela a seguir apresenta as variáveis que serão agrupadas, bem como seus valores após a realização de novos grupos.

Tabela 4: Distribuição dos valores da variável Quarto

Quarto	Frequência
1	8
2	33
3	472
4	4228
5	2875
6	863
7	206
8	96

Tabela 5: Distribuição dos valores agrupados da variável Quarto

Quarto	Frequência
1-2	41
3	472
4	4228
5+	4040

Tabela 6: Distribuição dos valores da variável Suíte

Suíte	Frequência
0	5
1	693
2	1470
3	1286
4	2852
5	2082
6	395
7	50
8	29

Tabela 7: Distribuição dos valores agrupados da variável Suíte

Suíte	Frequência
1-2	698
3	1286
4	2852
5+	2556

Tabela 8: Distribuição dos valores da variável Vaga

Vaga	Frequência
1	94
2	956
3	1207
4	3149
5	781
6	1549
7	212
8	745

Tabela 9: Distribuição dos valores agrupados da variável vaga

Vaga	Frequência
1-2	1050
3-4	4356
5+	3287

4.3 Transformação dos dados

Nessa etapa as variáveis são definidas como quantitativas ou qualitativas. Em caso de a variável independente ser qualitativa, ela deverá ser transformada em uma variável do tipo dummy e, caso apresente uma determinada característica (pertencer a uma região da cidade estudada, por exemplo), é atribuído valor 1 (um) à variável; e em caso negativo, é atribuído valor 0 (zero). As tabelas 10 e 11 ilustram a utilização de variáveis dummies em relação a variável "Bairro"e "Ano", as quais possuem 2 níveis.

Tabela 10: Variável Dummy criada para a variável Bairro

Bairro de Brasília	Variável Dummy
Lago Norte	1
Lago Sul(Não Lago Norte)	0

Tabela 11: Variável Dummy criada para a variável Ano

Ano	Variável Dummy
2020	1
2021(Não 2020)	0

No caso de algumas variáveis quantitativas, elas podem apresentar grande variação nos valores. Desse modo, devem ser feitas transformações para a limitação desses valores por meio de pesos e, assim, ser corrigido o problema de linearidade. Assim, com o intuito de obter resultados satisfatórios quanto à assimetria, foi utilizada a transformação logarítmica na variável resposta. A transformação logarítmica foi escolhida por ser sugerida pela norma e por proporcionar uma variância do erro do modelo próxima de zero, permitindo um melhor ajuste do modelo.

4.4 Modelo de Regressão linear

A princípio, no presente trabalho, os preços dos imóveis foram avaliados em função das variáveis quantitativas, tendo sido a variável bairro trabalhada posteriormente, por ser uma variável Dummy e demandar tratamento estatístico diferenciado.

Ao ajustar os modelos, é possível aumentar a probabilidade adicionando parâmetros, mas isso pode resultar em overfitting. Assim, o BIC tenta resolver esse problema introduzindo um termo de penalidade para o número de parâmetros no modelo. Nesse sentido, temos o seguinte resultado:

Tabela 12: Critério BIC para o modelo de Regressão com a variável Bairro

Step	Variável	BIC
0	Intercepto	-5662,66
1	Suítes	-7334,64
2	Bairro	-7870,31
3	Área	-8206,02
4	Ano	-8334,96
5	Vaga	-8411,70*
* Valor ótimo para o critério BIC		

A seleção passo a passo parou porque adicionar ou remover um efeito não melhora o critério BIC. Ao se calcular os coefientes e construir o modelo ajustado a partir das observações da base de treinamento se obtém o seguinte resultado:

Tabela 13: Tabela do Modelo de Regressão com variável Bairro

Efeito	Parâmetro	Variável	Estatística t	Estimativa	Erro Padrão	Pr $\mid t \mid$
Intercepto	Intercepto		816,94	15,43	0,02	0
áreas	áreas	V1	18,72	$0{,}02 \cdot 10^{-2}$	$<,0001$	$<,0001$
ano	ano 2020	V2	-11,91	-0,15	0,01	$<,0001$
bairro	Lago Norte	V3	-26,29	-0,35	0,013	$<,0001$
suíte	suíte 1-2	V4_1	-35,24	-0,63	0,02	$<,0001$
suíte	suíte 3	V4_2	-23,73	-0,48	0,02	$<,0001$
suíte	suíte 4	V4_3	-11,76	-0,19	0,02	$<,0001$
vaga	vaga 1-2	V5_1	-8,26	-0,17	0,02	$<,0001$
vaga	vaga 3-4	V5_2	-7,83	-0,11	0,01	$<,0001$

Analisando a Tabela 13 e considerando um nível de significância de 0,05, o modelo seria definido da seguinte forma:

$$log(y) = 15,43 + 0,02 \cdot 10^{-2} V1 - 0,15 V2 - 0,35 V3 - 0,63 V4_1$$

$$-0,48V4_2 - 0,19V4_3 - 0,17V5_1 - 0,11V5_2$$

Nesse sentido, tendo como referência a variável **suíte**, caso um imóvel em questão tenha 3 suítes, o valor do mesmo é 16% superior a outra residência com as mesmas características porem com o número de suítes entre 1 e 2. O mesmo ocorre para o número de vagas, caso um imóvel tenha mais de 4 vagas, o seu valor é 11,6% superior a outra residência a qual possua as mesmas características porém, com o número de vagas entre 3 e 4.

4.4.1 Modelo com a variável cluster para quadra

A utilização da variável quadra como Dummy prejudica a eficiência do modelo, visto que o objetivo do estudo é desenvolver um modelo de regressão linear com o melhor desempenho possível e utilizando apenas variáveis relevantes. Nesse sentido, foram criados 4 grupos os quais, de certa forma, representam a similaridade social e de infra-estrutura dos bairros que compõem o grupo. Tais grupos foram realizados por meio da média de valor encontrada para a quadra referente. Nesse sentido, temos o seguinte resultado:

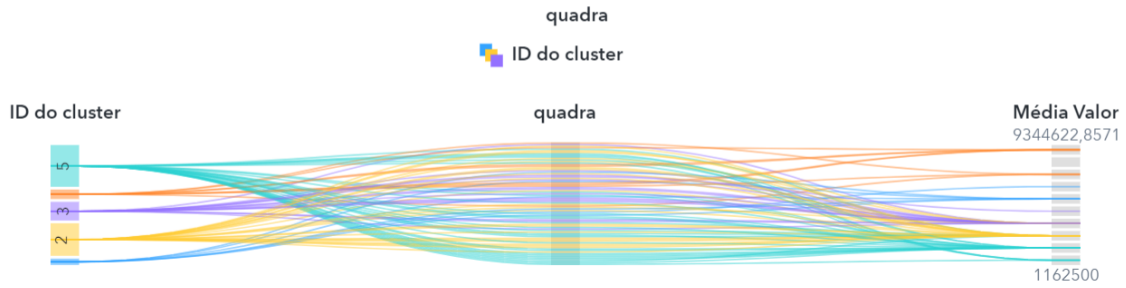


Figura 13: Relacionamento da variável cluster com a Quadra do imóvel

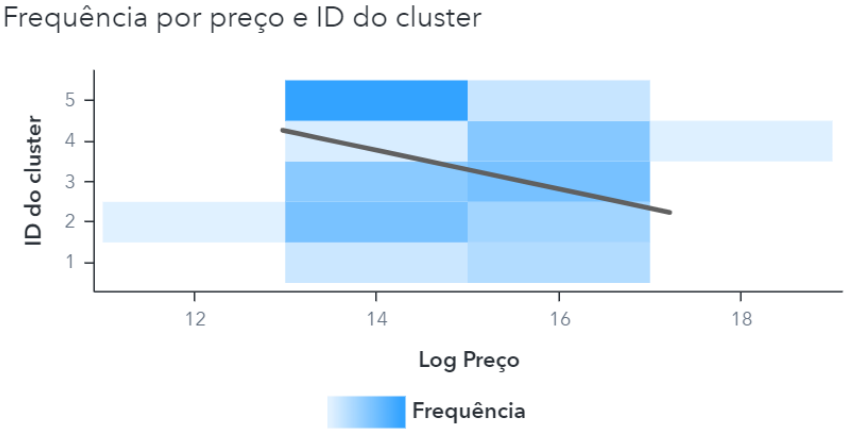


Figura 14: Gráfico de calor da Variável cluster para Quadra por $log(preco)$

Tabela 14: Quadras inseridas na variável cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
SHIN QL 7	SHIS QI 3	SMLN ML 7	SHIS QL 10	SHIN QI 1
SHIN QL 9	SHIN QI 8	SHIS QI 7	SHIS QL 6	SHIS QI 29
SHIS QI 11	SMLN MI 3	SHIN QL 3	SHIS QI 15	SMLN ML 3
SHIS QI 26	SHIN QI 4	SHIS QL 26	SMLN ML 4	SHIS QI 27
SHIS QL 14	SHIN QI 3	SHIS QI 23	SMLN ML 9	SHIN QI 9
SHIS QL 24	SHIS QI 1	SHIS QI 5	SHIS QI 9	SMLN MI 7
	SMLN MI 6	SHIS QI 25	SHIS QL 8	SHIN QI 15
	SHIN QI 7	SHIS QL 16	SHIS QL 12	SHIN CA 6
	SHIN QI 6	SHIN QL 13		SHIN QL 1
	SHIS QL 18	SHIN QL 4		SMLN MI 13
	SHIN QI 14	SHIS QI 28		SMLN MI 9
	SHIS QI 13	SHIS QL 20		SHIN QI 10
	SHIS QL 28	SHIS QL 2		SHIN QL 5
	SHIN QL 11	SHIS QI 21		SMLN MI 4
	SMLN ML 6			SHIN QL 2
	SHIN QL 16			SHIN CA 2
	SMLN ML 12			SHIN QL 10
	SHIN QL 15			SHIS QI 17
	SHIS QL 22			SHIN CA 10
	SHIN QI 2			SHIN QI 12
	SHIS QL 4			SHIN QI 13
	SMLN ML 13			SHIN QI 16
	SHIS QI 16			SHIN QL 12
				SHIN QI 5
				SHIN QL 6
				SMLN MI 8
				SHIN QL 14
				SHIN QI 11
				SHIN QL 8

Por fim, temos o seguinte resultado:

Tabela 15: Critério BIC para o modelo de Regressão com a variável Cluster

Step	Variável	BIC
0	Intercepto	-5662,66
1	ID do Cluster	-8159,26
2	Suíte	-9592,57
3	Área	-9795,68
4	Ano	-9885,63
5	Vaga	-9974,44*
* Valor ótimo para o critério BIC		

Tabela 16: Tabela do Modelo de Regressão com variável cluster

Efeito	Parâmetro	Variável	Estatística t	Estimativa	Erro Padrão	Pr $\mid t \mid$
Intercepto	Intercepto		833,13	14,94	0,01	0
areas	areas	V1	14,65	$0{,}02 \cdot 10^{-4}$	$<, 0001$	$<, 0001$
ano	ano 2020	V2	-10,11	-0,11	0,01	$<, 0001$
ID do cluster	ID 1	V3_1	27,96	0,58	0,02	$<, 0001$
ID do cluster	ID 2	V3_2	15,86	0,24	0,02	$<, 0001$
ID do cluster	ID 3	V3_3	28,04	0,41	0,01	$<, 0001$
ID do cluster	ID 4	V3_4	50,00	0,91	0,02	$<, 0001$
suíte	suíte 1-2	V4_1	-32,8718	-0,02	0,52	$<, 0001$
suíte	suíte 3	V4_2	-23,4742	-0,02	0,41	$<, 0001$
suíte	suíte 4	V4_3	-11,712	-0,16	0,01	$<, 0001$
vaga	vaga 1-2	V5_1	8,916233	-0,16	0,02	$<, 0001$
vaga	vaga 3-4	V5_2	-8,15912	-0,09	0,01	$<, 0001$

Analisando a Tabela 16 e considerando um nível de significância de 0,05, o modelo seria definido da seguinte forma:

$$log(y) = 14,94 + 0,02 \cdot 10^{-4} V1 - 0,115 V2 + 0,58 V3_1 + 0,24 V3_2$$

$$+0,41 V3_3 + 0,91 V3_4 - 0,52 V4_1 - 0,41 V4_2 - 0,16 V4_3 - 0,16 V5_1 - 0,09 V5_2$$

Nesse sentido, tendo como referência a variável **Cluster**, caso um imóvel em questão esteja setorizado na **Cluster 1**, o valor do mesmo é 18% superior a outra residência com as mesmas características porem localizado na **Cluster 3**.

4.5 Modelo Linear Generalizado

A ideia básica sobre os MLGs consiste em abrir o leque de opções para a distribuição da variável resposta, permitindo que ela pertença à família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear. Sobre a população de imóveis de interesse, foram estimados cinco modelos lineares generalizados (MLGs), um para cada região do DF. O interesse é avaliar de que forma o número de quartos, o número de suítes, o número de garagens, a área útil e os bairros, quando possível, impactam o valor dos imóveis. Os modelos estimados assumem distribuição gama para a variável resposta (valor do imóvel), já que essa distribuição é indicada em casos de respostas positivas assimétricas. Pode-se mostrar também que à medida que o parâmetro de dispersão cresce, a distribuição gama fica mais simétrica em torno da média, e se aproxima de uma distribuição normal. Portanto, a distribuição gama também é atrativa em casos de respostas simétricas. As funções de ligação mais usadas no caso gama são a identidade, logarítmica e recíproca. Nesse estudo usaremos a logarítmica, pois possibilita o desenvolvimento de experimentos ortogonais como são bem conhecidos em modelos de regressão normal linear, além de possibilitar interpretações interessantes sobre o impacto em termos percentuais das variáveis explicativas sobre a variável resposta Nesse sentido, o BIC, o qual tem como pressuposto a existência de um "modelo verdadeiro" o qual descreve a relação entre a variável dependente e as diversas variáveis explanatórias entre os diversos modelos sob seleção, resultou no seguinte resultado:

Tabela 17: Critério BIC para o modelo de regressão logística com a variável Bairro

Step	Variável	BIC
0	Intercepto	170464,37
1	Suítes	168935,29
2	Bairro	168075,65
3	Área	167375,48
4	Ano	167218,52
5	Vaga	167125,32
6	Quarto	167114,14*
* Valor ótimo para o critério BIC		

O modelo na etapa 6 é selecionado onde BIC é 167114,14. Por fim, temos as seguintes estimativas de parâmetros:

Tabela 18: Tabela do Modelo de regressão logística com variável Bairro

Efeito	Parâmetro	Variável	Estatística t	Estimativa	Erro Padrão	Pr | t |
Intercepto	Intercepto		717,42	14,81	$0{,}12 \cdot 10^{-2}$	$<, 0001$
áreas	áreas	V1	26,64	$0{,}16 \cdot 10^{-4}$	$<, 0001$	$<, 0001$
ano	ano 2020	V2	-12,03	-0,01	$0{,}08 \cdot 10^{-2}$	$<, 0001$
bairro	Lago Norte	V3	-26,43	-0,02	$0{,}09 \cdot 10^{-2}$	$<, 0001$
suíte	suíte 1-2	V4_1	-15,10	-0,03	$0{,}11 \cdot 10^{-2}$	$<, 0001$
suíte	suíte 3	V4_2	-23,73	-0,02	$0{,}13 \cdot 10^{-2}$	$<, 0001$
suíte	suíte 5-	V4_4	11,50	0,01	$0{,}11 \cdot 10^{-2}$	$<, 0001$
vaga	vaga 1-2	V5_1	-3,49	$-0{,}46 \cdot 10^{-2}$	$0{,}13 \cdot 10^{-2}$	$<, 0001$
vaga	vaga 5-	V5_3	7,74	0,01	$0{,}09 \cdot 10^{-2}$	$<, 0001$

Analisando a Tabela 18 e considerando um nível de significância de 0,05, o modelo seria definido da seguinte forma:

$$log(y) = 14,81 + 0,16 \cdot 10^{-4} V1 - 0,01 V2 - 0,02 V3 - 0,03 V4_1$$

$$-0,02 V4_2 + 0,01 V4_4 - 0,46 \cdot 10^{-2} V5_1 + 0,09 \cdot 10^{-2} V5_3$$

Nesse sentido, tendo como referência a variável **suíte**, caso um imóvel em questão tenha 3 suítes, o valor do mesmo é 30% superior a outra residência com as mesmas características porem com o número de suítes entre 1 e 2.

4.6 Modelo com a variável cluster para quadra

Utilizando a variável Cluster criada com o intuito de alocar a variável bairro em 4 grupos os quais possuem características semelhantes entre si, foi-se observado o seguinte resultado :

Tabela 19: Critério BIC para o modelo de regressão logística com a variável Cluster

Step	Variável	BIC
0	Intercepto	9376,30
1	ID do Cluster	7374,24
2	Suíte	5618,26
3	Área	5232,47
4	Ano	5125,03
5	Vaga	5037,55
6	Quarto	5033,22*
* Valor ótimo para o critério BIC		

Tabela 20: Tabela do Modelo de regressão logística com variável cluster

Efeito	Parametro	Variável	Estatística t	Estimativa	Erro Padrão	Pr $\mid t \mid$
Intercepto	Intercepto		834,76	14,69	0,01	$<,0001$
areas	areas	V1	49,84	$0,11 \cdot 10^{-2}$	$0,02 \cdot 10^{-2}$	$<,0001$
ano	ano 2020	V2	-10,52	-0,01	$0,07 \cdot 10^{-2}$	$<,0001$
ID do cluster	ID 1	V3_1	28,19	0,04	$0,14 \cdot 10^{-2}$	$<,0001$
ID do cluster	ID 2	V3_2	16,10	0,02	$0,10 \cdot 10^{-2}$	$<,0001$
ID do cluster	ID 3	V3_3	28,22	0,03	$0,10 \cdot 10^{-2}$	$<,0001$
ID do cluster	ID 4	V3_4	49,84	0,06	$0,12 \cdot 10^{-2}$	$<,0001$
suíte	suíte 1-2	V4_1	-23,14	-0,02		$<,0001$
suíte	suíte 3	V4_2	-12,61	-0,02	$0,12 \cdot 10^{-2}$	$<,0001$
suíte	suíte 5	V4_4	9,55	0,01	$0,12 \cdot 10^{-2}$	$<,0001$
vaga	vaga 1-2	V5_1	-3,81	$-0,44 \cdot 10^{-2}$	$0,11 \cdot 10^{-2}$	$<,0001$
vaga	vaga 5	V5_3	7,97	$0,61 \cdot 10^{-2}$	$0,08 \cdot 10^{-2}$	$<,0001$
quarto	quarto 1-2	V6_1	2,10	-0,02	$<,0001$	$<,0001$
quarto	quarto 3	V6_2	1,32	-0,01	$<,0001$	$<,0001$
quarto	quarto 5	V6_4	0,83	$-0,08 \cdot 10^{-2}$	$<,0001$	0,04

Analisando a Tabela 20 e considerando um nível de significância de 0,05, o modelo seria definido da seguinte forma:

$$log(y) = 14,69 + 0,11 \cdot 10^{-2}V1 - 0,01V2 + 0,04V3_1 + 0,02V3_2 + 0,03V3_3 + 0,06V3_4 -$$

$$0,02V4_1 - 0,41V4_2 + 0,01V4_4 - 0,44 \cdot 10^{-2}V5_1 + 0,61 \cdot 10^{-2}V5_3 - 0,02V6_1 - 0,01V6_2 - 0,08 \cdot 10^{-2}V6_4$$

Nesse sentido, tendo como referência a variável **Cluster**, caso um imóvel em questão esteja setorizado na **Cluster 2**, o valor do mesmo é 20% superior a outra residência com as mesmas características porem localizado na **Cluster 1**.

4.7 Validação dos modelos

Dessa forma, após realizar a construção dos modelos por meio da plataforma apresentada, é possível observar uma comparação entre eles, bem como uma análise dos dados do modelo final, os quais foram divididos da seguinte maneira:

- Dados de treinamento: usado para treinar o modelo;
- Dados de validação: usado para comparação de diferentes modelos e hiper parâmetros;
- Dados de teste: usado para comprovar que aquele modelo realmente funciona. São dados ignorados no treinamento e no processo de escolha de hiper parâmetros.

O modelo de regressão clássico pressupõe que a variável resposta seja simétrica e homocedástica. Porém, em muitas situações esses pressupostos não são alcançados e precisa-se de uma abordagem mais flexível que alcance dados de natureza contínua com comportamento positivo assimétrico. Com efeito, o Modelo de regressão logística, por ser versátil, permite que a variável resposta se adéque a esse comportamento. Nesse sentido, tem-se na Tabela 21 o erro quadrático médio (MSE), o qual fornece a média de diferença quadrática entre a predição do modelo e o valor de destino:

Tabela 21: Erro Quadrático Médio para os modelos

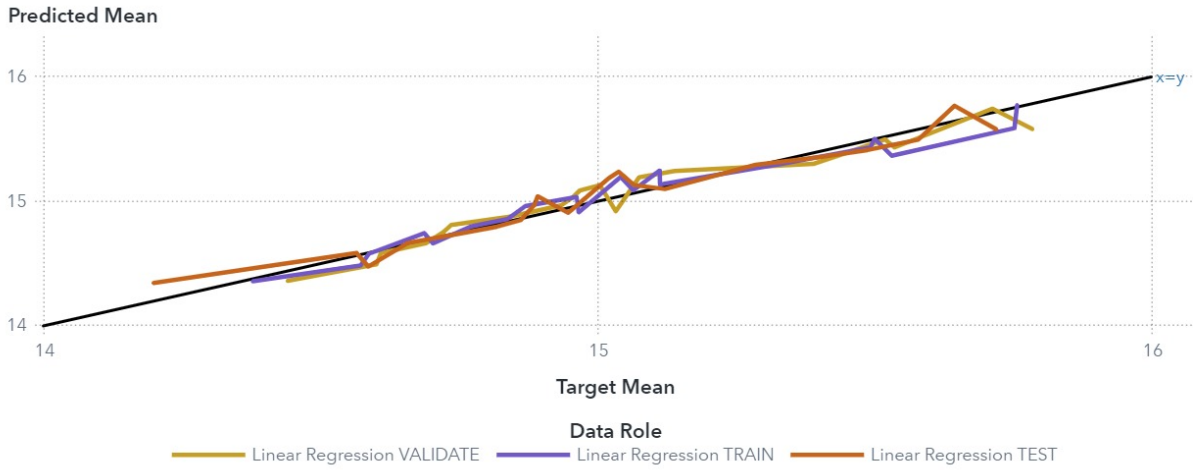
Modelo	MSE
Regressão Linear Múltipla	0,183
Regressão Linear Múltipla com variável Cluster	0,143
Regressão logística	0,164
Regressão logística com variável Cluster	0,142

Assim, analisando os resultados apresentados, segue que o modelo o qual melhor representou os dados, bem como explicou a variável resposta **preço**, foi o modelo de regressão logística com a utilização da variável Cluster.

4.7.1 Modelo de Regressão Linear Múltipla

Para o modelo de regressão linear múltipla tem-se o seguinte resultado:

Figura 15: Resultado para o modelo de Regressão sem a varíavel Cluster

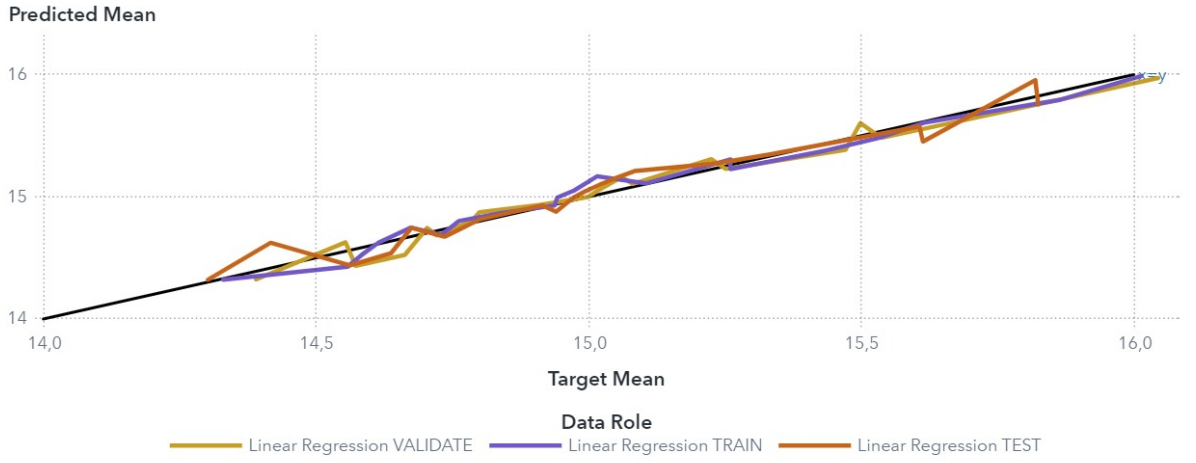


Para este gráfico, bem como para os demais, cada partição dos dados é classificada pelo alvo previsto >preço. para o alvo real, em ordem decrescente. Os dados são então divididos em 20 quantis (semi decis, com 5% dos dados em cada um), e a média do alvo previsto e do alvo real são calculados e plotados para cada quantil (profundidade em incrementos de 5).

A maior diferença entre as médias reais e previstas do alvo é 0,204 e ocorre para a partição de validação em profundidade 10. A metodologia apresentada para a realização das previsões de valores de imóveis se mostrou factível e precisa.

Ademais, segue o resultado para tal modelo com a inserção da variável Cluster:

Figura 16: Resultado para o modelo de Regressão com a varíavel Cluster

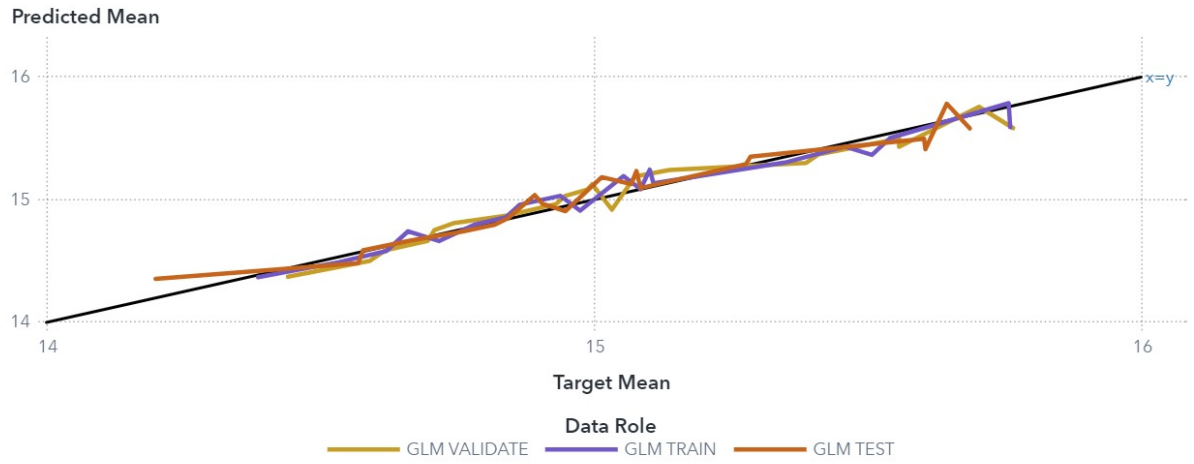


A maior diferença entre as médias reais e previstas do alvo é 0,192. A metodologia apresentada para a realização das previsões de valores de imóveis se mostrou factível e precisa.

4.7.2 Modelo Linear Generalizado

Para o modelo de regressãp logística tem-se o seguinte resultado:

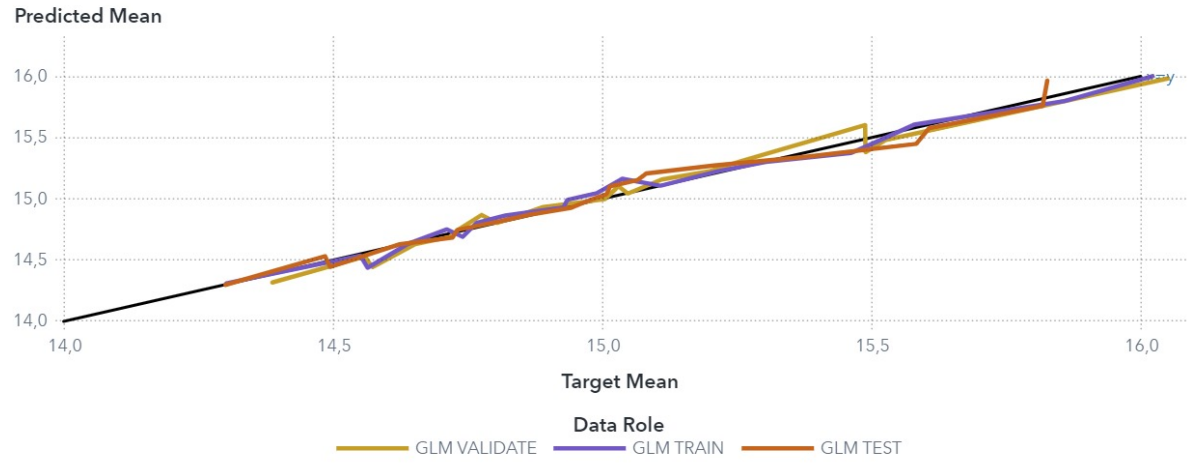
Figura 17: Resultado para o modelo de regressão logística sem a variável Cluster



A maior diferença entre as médias reais e previstas do alvo é de 0,193.

Ademais, segue o resultado para tal modelo com a inserção da variável Cluster:

Figura 18: Resultado para o modelo de regressão logística com a variável Cluster



A maior diferença entre as médias reais e previstas do alvo é 0,178. A metodologia apresentada para a realização das previsões de valores de imóveis se mostrou factível e precisa.

5 Conclusão

Dispondo de um banco de dados com amostras variadas com as características dos imóveis é possível representar o valor de mercado de um imóvel por meio de uma equação, Dentre as características foi possível utilizar as variáveis quantitativas e qualitativas das amostras.

Apesar de muitas variáveis independentes que apresentaram boa correlação com a variável dependente não terem entrado na equação, não significa que estas não são importantes para a formação do valor do imóvel em outros métodos de avaliação. Elas não entraram na equação final por apresentarem forte correlação com outras varáveis independentes presentes na equação, o que provocaria uma redundância dentro do modelo proposto. Portanto, não se pode afirmar que apenas as variáveis presentes no modelo são formadoras de valor de apartamentos residenciais nos bairros do Lago Norte e Lago Sul; estas são importantes para este modelo de avaliação específico. Com o intuito de agregar informação ao modelo, características adicionais podem ser estudadas para assim, realizar um investigar se essas influenciam no modelo, tais quais: presença de piscina na residência, reforma recente na casa, presença de parques e supermercados perto do local. Estas não forma acrescentadas devido ao banco de dados não possuir tais instruções.

Sendo assim, os valores apresentados referente a maior diferença entre as médias reais e previstas do alvo podem ser explicadas por tal fator. O modelo apresentado se trata de um estudo referente ao valor médio observado entre os imóveis utilizando as informações disponibilizadas. Contudo, é notório o fator de que, tais modelos não conseguem prever exatamente o valor do imóvel pelo fato de características adicionais influenciarem no preço. Além disso, faz-se relevante o fato que, o preço utilizado se trata dos valores disponibilizados no site **61 imóveis**, não necessariamente é o valor de venda do mesmo. Contudo, os modelos apresentados obtiveram ótimo desempenho de previsão de valores, sendo o modelo de regressão logística com a utilização da variável Cluster o que obteve melhor desempenho.

Por fim, a administração do Distrito Federal pode fazer uso da ferramenta como auxílio no cálculo dos tributos referentes a esse tipo de imóvel. As empresas construtoras e incorporadoras também podem utilizar a equação como apoio no momento de definição do preço dos apartamentos a serem lançados e, também, utilizar a equação para definição de alguma característica do seu novo empreendimento ao estipular variáveis.

Referências

ARRAES, R. A.; FILHO, E. d. S. Externalidades e formação de preços no mercado imobiliário urbano brasileiro: um estudo de caso. *Economia aplicada*, SciELO Brasil, v. 12, p. 289–319, 2008.

ASEVEDO, F. R. de. Abordagem linear generalizada para estimar perdas não técnicas de energia elétrica. Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, 2011.

BRITO, M.; RODRIGUES, A. O. Gt. 11-utopia e distopia no pensamento político moderno o sonho da casa prÓpria: Entre a utopia e a distopia. 2014.

CBIC. *Setor imobiliário tem retorno de intenção de compra para patamar pré-pandemia*. [S.l.]: Recuperado de https://cbic.org.br/industriaimobiliaria/2020/09/04/setor-imobiliario-tem-retorno-de-intencao-de-compra-para-patamar-pre-pandemia-2/, 2020.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP*, 2008.

HELENE, O. *Metodos dos Minimos Quadrados*. [S.l.]: Editora Livraria da Física, 2006.

HENRIQUES, C. Análise de regressão linear simples e múltipla. *Departamento de Matemática. Escola Superior de Tecnologia de Viseu. Portugal*, 2011.

JOVEMPAN. *Queda na Selic aquece mercado imobiliário e especialistas alertam: 'É o momento para comprar imóveis'*. [S.l.]: Recuperado de https://jovempan.com.br/noticias/economia/queda-na-selic-aquece-mercado-imobiliario-e-especialistas-alertam-e-o-momento-para-comprar-imoveis.html, 2020.

KASSAMBARA, A. *Practical guide to cluster analysis in R: Unsupervised machine learning*. [S.l.]: Sthda, 2017. v. 1.

MATOS, D.; BARTKIW, P. I. N. Introdução ao mercado imobiliário. *Curitiba: Instituto Federal de Educação, Ciência e Tecnologia–Paraná–Educação a distância*, 2013.

MISSIO, F. M.; JACOBI, L. F. Variáveis dummy: especificações de modelos com parâmetros variáveis. *Ciência e Natura*, p. 111–135, 2007.