UNIVERSIDADE FEDERAL DE MINAS GERAIS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Felipe Francisco Ferreira Alves do Carmo

PRECIFICAÇÃO DE IMÓVEIS NA CIDADE DE BELO HORIZONTE: UMA ANÁLISE PREDITIVA USANDO REGRESSÃO LINEAR E RANDOM FOREST

Belo Horizonte
2025

Felipe Francisco Ferreira Alves do Carmo

PRECIFICAÇÃO DE IMÓVEIS NA CIDADE DE BELO HORIZONTE: UMA ANÁLISE PREDITIVA USANDO REGRESSÃO LINEAR E RANDOM FOREST

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Orientador: Prof. Dr. Guilherme Lopes de Oliveira

Belo Horizonte

2025

Carmo, Felipe Francisco Ferreira Alves do.

C287p Precificação de imóveis na cidade de Belo Horizonte: [recurso eletrônico] uma análise preditiva usando regressão linear e random forest / Felipe Francisco Ferreira Alves do Carmo – 2025.
1 recurso online (53 f. il., color.) : pdf.

Orientador: Guilherme Lopes de Oliveira.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 53.

1. Estatística. 2. Análise de regressão. 3. Mercado imobiliário – Belo Horizonte (MG). 4. Mercado imobiliário – Controle de preços. 5. Árvores de decisão. 6. Floresta aleatória I. Oliveira, Guilherme Lopes de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação
Caixa Postal 702
31270-901 Belo Horizonte- MG – Brasil

Telefone (31) 3409-5923
Fax (31) 3499-5924
E-mail: pgest@ufmg.br
WEB: http://www.est.ufmg.br/posgrad/

ATA DO 346ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE FELIPE FRANCISCO FERREIRA ALVES DO CARMO.

Aos onze dias do mês de abril de 2025, às 09:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística Computacional Aplicada, para julgar a apresentação do trabalho de fim de curso do aluno **Felipe Francisco Ferreira Alves do Carmo**, intitulado: "*Precificação de imóveis na cidade de Belo Horizonte: uma análise preditiva usando regressão linear e random forest"*, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 11 de abril de 2025.

gov.br Documento assinado digitalmente
GUILHERME LOPES DE OLIVEIRA
Data: 11/04/2025 10:38:18-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Lopes de Oliveira (orientador)
DECOM/CEFET-MG

gov.br Documento assinado digitalmente
GUILHERME AUGUSTO VELOSO
Data: 11/04/2025 15:25:55-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Augusto Veloso
GET/UFF

RESUMO

O mercado imobiliário exerce grande importância e relevância para a economia dos municípios brasileiros. O presente trabalho busca aplicar modelos preditivos para a precificação de imóveis do tipo casa e apartamento em Belo Horizonte, Minas Gerais, Brasil. Através da base de dados pública relativa ao recolhimento do Imposto sobre Transmissão de Bens Imóveis (ITBI), disponível no portal de dados abertos da Prefeitura de Belo Horizonte, foi realizada uma análise desses dados utilizando duas técnicas: regressão linear e Random Forest. Variáveis como a regional do município, tipo de acabamento do imóvel e idade da construção foram consideradas como possíveis preditores. Os resultados indicaram que o modelo Random Forest apresentou 82% de precisão na precificação de um imóvel na cidade de Belo Horizonte com base em tais fatores. A pesquisa contribui para exemplificar como métodos estatísticos podem ser utilizados para guiar e avaliar uma precificação imobiliária mais realista e alinhada aos preços praticados no mercado imobiliário do município.

Palavras-chave: precificação; mercado imobiliário; predição; random forest.

ABSTRACT

The real estate market plays a significant role in the economy of Brazilian municipalities. This study aims to apply predictive models for pricing residential properties, such as houses and apartments, in Belo Horizonte, Minas Gerais, Brazil. Using a public database related to the collection of the Buildings Transfer Tax (ITBI), available on the open data portal of the Belo Horizonte city hall, an analysis of these data was conducted using two techniques: linear regression and Random Forest. Variables such as the city's region, property finishing type, and building age were considered as potential predictors. The results indicated that the Random Forest model achieved 82% accuracy in predicting property prices in Belo Horizonte based on these factors. This research contributes to illustrating how statistical methods can be used to guide and assess a more realistic property valuation, aligning with market prices in the municipality's real estate sector.

Keywords: pricing; real estate market; predictors; random forest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O mercado imobiliário de Belo Horizonte desempenha um papel crucial na economia local, de acordo com *Diário do Comércio* (2024), entre setembro de 2023 e setembro de 2024, o setor alcançou um valor geral de vendas de R$5,14 bilhões em Belo Horizonte, capital do Estado de Minas Gerais, Brasil. Isto evidencia como o setor imobiliário influencia diretamente o desenvolvimento econômico e urbano, gerando empregos diretos e indiretos em áreas como construção civil, arquitetura, venda de imóveis e financiamentos imobiliários. Além de ter impacto nas dinâmicas econômicas locais, o mercado financeiro também tem efeito direto na arrecadação fiscal.

Este relevante mercado tem sido notado em âmbitos acadêmicos. Estudos como o de Neri (2020), que estabeleceu um modelo preditivo para o preço de venda de apartamentos na cidade de Belo Horizonte utilizando Random Forest, e Paixão (2015), que analisou a valorização dos preços de imóveis de 1995 a 2003 utilizando o método dos preços hedônicos, que calcula índices de preços controlados pelas características dos bens, com base nos dados do Imposto de Transmissão de Bens Imóveis (ITBI), disponibilizados pela Secretaria de Fazenda do município de Belo Horizonte (SEFAZ/PBH), contribuem significativamente para a compreensão das variáveis que afetam o valor dos imóveis. O modelo de Neri (2020) obteve um desempenho robusto concatenando a simplicidade no uso do Random Forest, aliada à sua eficácia na manipulação de grandes conjuntos de variáveis e à boa capacidade de generalização, reforçando seu potencial como ferramenta de apoio à precificação de imóveis. Já Paixão (2015) demonstrou que área do imóvel, padrão de acabamento, distância ao centro e hierarquia socioespacial foram determinantes para a formação dos preços, assim como juros altos, baixo crescimento e queda no crédito habitacional mostraram que fatores macroeconômicos podem influenciar na precificação dos imóveis no período analisado.

Além disso, as transações imobiliárias não se limitam a influenciar o cotidiano dos cidadãos, mas também se tornam uma importante fonte de receita para a Prefeitura de Belo Horizonte, por meio do ITBI. Imposto este, pago no momento da

transferência de propriedades, que tem um papel fundamental no planejamento urbano da cidade, através do investimento em infraestrutura e serviços urbanos essenciais.

O registro de pagamento deste imposto disponibilizado no portal de dados abertos da prefeitura de Belo horizonte, possibilita uma análise para compreender a dinâmica das transações imobiliárias, oferecendo insights valiosos sobre o comportamento do mercado de venda de imóveis no município. Através dos dados do Portal de Dados Abertos da Prefeitura de Belo Horizonte (PBH), que abrange informações detalhadas sobre os lançamentos quitados de ITBI ao longo de mais de uma década, torna-se possível a realização de estudos com o intuito de predizer o valor de um imóvel baseado indicativos geográficos e estruturais dos imóveis presentes nessa base de dados pública. Este tipo de estudo pode ser interessante tanto para agentes do mercado imobiliário, como corretores e imobiliárias, ao facilitar seu trabalho de precificação de imóveis e aprimorar suas estratégias de negócio quanto para os cidadãos interessados em adquirir imóveis por um preço justo dado o contexto regional e estrutural associado ao bem

Neste contexto, este trabalho visa desenvolver modelos de predição de preço baseado em características geográficas como a regional da cidade de Belo Horizonte onde o imóvel se encontra, a área construída adquirida na compra do imóvel, o seu tipo construtivo podendo ser casa ou apartamento, o padrão de acabamento categorizado como muito simples, simples, médio, bom e muito bom, além da idade do imóvel no momento da compra, além de indicadores macroeconómicos como os índices de construção civil e variações de aluguéis residenciais da cidade de Belo Horizonte. Para este fim, serão consideradas transações de venda de imóveis residenciais registradas na base de dados pública da PBH entre janeiro de 2019 e maio de 2024.

Na sequência deste trabalho, a Seção 2 apresenta os dados e as metodologias utilizadas para alcançar os resultados exibidos na Seção 3. A Seção 4 encerra o estudo com algumas considerações finais.

2. MATERIAIS E MÉTODOS

2.1 Coleta e Definição das Bases de Dados

Para embasamento do trabalho, foi estipulado o uso de uma base de dados pública e acessível livremente. Para isso foi encontrada no portal de dados abertos da Prefeitura de Belo Horizonte uma base de dados relativa às transações de ITBI ocorridas no município.

Partindo do pressuposto que toda ação de transferência de imóveis, é declarada ao governo municipal para então ser registrada em cartório, a utilização dessa base serve como *proxy* para avaliar o comportamento dessas transações e por fim estimar um modelo que melhor se adeque a precificação de imóveis residenciais na cidade de Belo Horizonte.

A base de dados original extraída do portal de dados abertos da PBH, sendo denominada por: "01/1994 a 05/2024 - ITBI Relatórios", está disponibilizada em formato .csv no portal de dados abertos da Prefeitura de Belo Horizonte. Esta base possui 284.384 compras de imóveis disponíveis para análise, contendo dados de vendas realizadas no período entre janeiro de 1994 e maio de 2024. O dicionário dessa base de dados é apresentado na Tabela 1. Vale destacar que alguns filtros serão aplicados a essa base de dados inicial, como será descrito na sequência.

Tabela 1: Lista de variáveis iniciais selecionadas para o modelo.

Variável	Descrição	Tipo
Endereço	Endereço em que o imóvel está localizado.	Texto
Bairro	Bairro em que o imóvel está localizado.	Texto
Ano de Construção Unidade	Ano em que a construção do imóvel foi finalizada.	Texto
Área Terreno Total	Área de terreno do lote.	Real
Área Construída Adquirida	Resultado da aplicação do percentual adquirido pela Área de construção da unidade.	Real
Área Adquirida Unidades Somadas	Somatório das unidades nos casos em que temos vários tipos construtivos. Ocorre quando as unidades representam blocos distintos por não serem contíguos ou por constituírem economias diferentes.	Real

Padrão Acabamento Unidade	Padrão de Acabamento da unidade conforme a pontuação recebida das características construtivas.	Texto
Fração Ideal Adquirida	Incidência do percentual de aquisição sobre a fração ideal original.	Real
Tipo Construtivo Preponderante	Tipo construtivo com a maior Área.	Texto
Descrição Tipo Ocupação Unidade	Se o tipo de ocupação e residencial, não residencial ou territorial.	Texto
Valor Declarado	Valor de aquisição declarado pelo contribuinte.	Real
Valor Base Cálculo	Valor determinado pela administração tributaria através de avaliação com base nos elementos constantes do Cadastro Imobiliário ou o valor declarado pelo sujeito passivo, se este for maior.	Real
Zona Uso ITBI	Codificação do zoneamento urbano para o imóvel, atribuído pela Lei de Uso e Ocupação vigente na época da transação.	Texto
Data Quitação Transação	Data em que o ITBI foi quitado.	Data

Determinou-se o “Valor Declarado” como a variável resposta do estudo, por ser efetivamente o valor pago nas transações dos imóveis.

Com relação as demais variáveis, algumas como Área Construída Adquirida, Tipo Construtivo Preponderante e Padrão Acabamento Unidade, foram utilizadas como variáveis independentes para construção dos modelos preditivos.

As variáveis Área Terreno Total, Área Adquirida Unidades Somadas, Fração Ideal Adquirida foram dispensadas da análise, por não representarem características diretas para indicação no preço de um imóvel em si no escopo deste estudo, pois, por exemplo, a área total do condomínio (terreno ou construções somadas) onde está localizado um apartamento é menos relevante do que a área contruída da unidade, a qual dimensiona o tamanho real do imóvel adquirido pelo comprador. O Endereço também não foi considerado pois outras variáveis como Bairro já identificam de maneira abrangente a localização e apresenta um número menor de categorias.

Observando no dicionário dos dados disponibilizado pelo portal de dados abertos da prefeitura da cidade de Belo Horizonte, foi identificado que não há uma clara descrição do tipo da moeda negociada nas transações imobiliárias do período dos

dados. A fim de simplificar as análises sem a necessidade de realizar variações cambias entre as moedas utilizadas no Brasil desde a criação da cidade de Belo Horizonte em 1893, foi realizado um filtro a partir da variável Ano de Construção Unidade para que fossem considerados apenas imóveis construídos a partir de 1994, visto que esse foi o ano de troca da moeda brasileira, do cruzeiro para o real. O real é a moeda utilizada correntemente na negociação de imóveis na cidade de Belo Horizonte. Dessa forma, com dados compreendidos entre 1994 e 2024, trabalhou-se com uma faixa de imóveis com idade de no máximo 30 anos.

A variável Padrão Acabamento Unidade está identificada por códigos P1, P2, P3, P4 e P5, sendo fundamentados em regulamentações da Prefeitura de Belo Horizonte e seguem parâmetros definidos na legislação municipal, como a Lei nº 7.166 de 1996, que trata do Código de Edificações de Belo Horizonte, assim como as normas estabelecidas pelo Plano Diretor Municipal baseado na lei nº 11.181/2019. As definições destes códigos são dadas conforme Tabela 2. Nesta variável, todas as classificações foram utilizadas para definição do modelo preditivo e identificação de possíveis diferenciais de preço.

Tabela 2: Definição do tipo de acabamento dos imóveis.

Variável	**Descrição**	**Descrição**
P1	Padrão Muito Simples.	Refere-se a construções com materiais de menor qualidade, com acabamentos básicos e padrões construtivos voltados para o essencial. Normalmente usado em construções de interesse social ou moradias mais acessíveis.
P2	Padrão Simples.	Superior ao P1, mas ainda focado em acabamentos simples, com materiais que atendem ao básico sem grandes sofisticações. Geralmente inclui pisos cerâmicos simples e revestimentos básicos.
P3	Padrão Médio.	A qualidade dos materiais e acabamentos melhora, com um padrão mediano. Pode incluir pisos de cerâmica de melhor qualidade, revestimentos mais elaborados e melhores esquadrias.
P4	Padrão Bom.	Construções que já começam a incluir materiais de qualidade superior, com acabamentos detalhados, pisos de porcelanato, janelas de vidro temperado, entre outros detalhes que agregam valor.
P5	Padrão Muito Bom ou Alto.	O nível mais alto de acabamento, com materiais de alto padrão, como mármores, granitos, pisos de madeira nobre, infraestrutura de primeira linha, e detalhes sofisticados que elevam o valor da construção. Como essa variável caracteriza a qualidade dos imóveis e isso se relaciona diretamente com o valor do imóvel, todas as classificações

		dessa variável serão utilizadas na realização desse trabalho.

A variável Tipo Construtivo Preponderante diz respeito ao tipo de construção do imóvel. De acordo com o dicionário disponibilizado pela prefeitura, é possível identificar os códigos dessa variável da forma descrita na Tabela 3. Como o objetivo de estudo são apenas imóveis residenciais, com o auxílio adicional da variável Descrição Tipo Ocupação Unidade, foi realizado um filtro para considerar apenas o tipo construtivo apartamento (AP) e casa (CA) residenciais para a análise.

Tabela 3: Definição do tipo construtivo preponderante dos imóveis.

Variável	**Descrição**	**Descrição**
AC	Apartamento comercial.	Imóvel destinado a função diversa a habitação em construção destinada originalmente a habitação, em conjunto vertical residencial, multifamiliar, com áreas comuns e acesso ao logradouro por via comum, bem como as construções destinadas a atividade de apart-hotel.
AP	Apartamento.	Construção destinada à habitação multifamiliar em edificação vertical, com uma ou mais unidades por pavimento, com áreas comuns e acesso ao logradouro por via comum.
BA	Barracão.	Construção destinada à habitação que seja igual ou menor que 60m2 por unidade, podendo existir mais de um pavimento.
BC	Barracão comercial.	Construção destinada a função diversa a habitação que seja igual ou menor que 60m2 por unidade, podendo existir mais de um pavimento.
CA	Casa.	Construção destinada à habitação que tenha mais de 60 m², ou construções, de ocupação residencial, que não se enquadrem nos demais tipos construtivos residenciais descritos. As casas geminadas não serão definidas com o tipo construtivo Apartamento (AP) e sim, com o tipo construtivo Casa (CA), ainda que definidas como apartamento na Convenção de Condomínio registada em Serviço de Registro de Imóveis, tendo em vista a situação fática do imóvel.
CC	Casa comercial.	Construção destinada a função diversa a habitação que tenha mais de 60 m².
GP	Galpão.	Construção com um pavimento destinada a fins industriais, depósitos, oficinas, salão de produção, postos de gasolina, estacionamento de veículos ou outras prestações de serviço, admitindo a existência de mezanino ou jirau, de grandes vãos, com pé direito em torno de 5 m, ou vãos e

		pé-direito menores, desde que abertos ou com meia parede e não se enquadrem nas construções destinadas a atividades descritas na definição de LJ.
LJ	Loja.	Imóvel não residencial de rua ou localizado em centros comerciais destinados a exposição e venda de mercadorias de fabricação própria ou de terceiros, incluindo as construções destinadas a atividades como cinemas, teatros, templos religiosos, hotéis, motéis, mercados, supermercados, hipermercados, instituições financeiras, clubes esportivos e sociais, colégios, creches, guaritas de estacionamento, hospitais, clínicas e similares; ou construções, de ocupação não residencial, que não se enquadrem nos tipos construtivos não residenciais.
LV	Lote vago.	Terreno que não possui nenhum tipo de construção ou edificação.
SL	Sala.	Unidade não residencial destinada à prestação de serviços, em conjunto vertical com áreas e entradas comuns, sem acesso direto ao nível do logradouro.
VC	Vaga de garagem comercial.	Espaço físico demarcado destinado a permanência de veículos quando se construir em unidade autônoma condominial de uso não residencial ou edifícios garagem.
VR	Vaga de garagem residencial.	Espaço físico demarcado destinado a permanência de veículos quando se constituir em unidade autônoma condominial de uso residencial.
VV	Vaga de garagem uso misto.	Aplicava-se às vagas de garagem que tinham uso misto (comercial e residencial). Atualmente não é mais utilizada. Porém, pode vir a constar na transação de um imóvel antigo.

A variável Zona Uso ITBI é mais uma das variáveis que é descrita por códigos, sendo determinados pelo Plano Diretor definido pela Lei nº 11.181/2019 e pela Lei de Uso e Ocupação do Solo de Belo Horizonte atualizada recentemente pela Lei Complementar nº 240/2022. Cada zona tem regras específicas para definir o que pode ou não ser construído, os tipos de atividade permitidos, além de definir parâmetros urbanísticos como coeficiente de aproveitamento, taxa de ocupação e gabarito (número de andares permitido). Os códigos são descritos e exemplificados na Tabela 4. Como o estudo visa propor uma precificação de imóveis residenciais, baseado na Tabela 4, as zonas de Adensamento Restrito 2, Zona Central de Belo

Horizonte, Zona de Adensamento e Zona de Adensamento Preferencial foram filtradas por serem áreas destinadas as construções residenciais no município.

Tabela 4: Descrição das zonas de classificação dos imóveis para cálculo do ITBI.

Variável	Descrição	Descrição
ZEENG	Zona Especial de Interesse Econômico de Novo Gameleira	Áreas destinadas ao desenvolvimento econômico, com ênfase em atividades comerciais e industriais.
ZEPIL	Zona Especial de Interesse de Proteção da Paisagem e Identidade Local	Áreas com características de interesse especial para preservação da paisagem ou da identidade local, normalmente em locais com relevância histórica, cultural ou paisagística.
ZEJAT	Zona Especial de Interesse de Justiça Ambiental e Territorial	Regiões voltadas para a implementação de políticas públicas voltadas à justiça ambiental, geralmente áreas de vulnerabilidade social e ambiental.
ZEIS1	Zonas Especiais de Interesse Social 1	Áreas urbanas que visam à regularização fundiária de assentamentos de baixa renda, com prioridade para habitação social.
ZEIS2	Zonas Especiais de Interesse Social 2	Áreas voltadas para habitação social, com regras menos rígidas para ocupação e construção.
ZEIS3	Zonas Especiais de Interesse Social 3	Áreas destinadas a habitação de interesse social, mas que exigem uma urbanização mais consolidada e integrada ao tecido urbano.
ZPAM	Zona de Proteção Ambiental	Áreas destinadas à preservação de recursos naturais, como matas e cursos d'água. A ocupação é bastante restrita e voltada à proteção ambiental.
ZESFR.	Zona Especial de Sítios de Relevância Histórico-Cultural e Ambiental	Áreas que apresentam relevância histórica, cultural e ambiental, com regras específicas para a preservação dessas características.
ZCVN:	Zona de Consolidação e Valorização da Natureza	Áreas em que se prioriza a valorização e a conservação de aspectos naturais e paisagísticos.
ZP1	Zonas de Proteção 1	Áreas com alta restrição de uso, voltadas para a preservação total de recursos naturais e culturais.
ZP2	Zonas de Proteção 2	Áreas com restrições moderadas, onde há controle de ocupação e preservação ambiental.
ZP3	Zonas de Proteção 3	Áreas de transição entre áreas de preservação e de uso urbano, com regras intermediárias de ocupação.
ZCBA	Zona de Comércio e Serviços de Bairro	Áreas com permissão para atividades comerciais e de serviços, voltadas principalmente para o atendimento às

		demandas locais.
ZE	Zona Especial	Regiões que possuem características especiais ou que demandam regras urbanísticas diferenciadas para atender a contextos específicos de preservação, desenvolvimento ou revitalização.
ZAR1	Zonas de Adensamento Restrito 1	Áreas residenciais com baixo adensamento e com restrições para o aumento de densidade populacional.
ZAR2	Zonas de Adensamento Restrito 2	Similar à ZAR1, mas com regras ligeiramente mais permissivas para o adensamento e ocupação.
ZHIP	Zona de Habitação de Interesse Popular	Áreas destinadas à habitação de interesse popular, com incentivo à construção de moradias de baixo custo para famílias de baixa renda.
ZCBH	Zona Central de Belo Horizonte	A área central da cidade, onde há uma maior permissividade para o uso misto (residencial, comercial, institucional), buscando a revitalização e ocupação intensa dessa zona.
ZA	Zona de Adensamento	Área destinada ao adensamento urbano, com maior permissividade para construções residenciais e comerciais, incentivando a ocupação vertical.
ZAP	Zona de Adensamento Preferencial	Áreas específicas com grande potencial para adensamento e verticalização, buscando atender ao crescimento populacional e econômico da cidade.

A variável Data Quitação Transação foi tratada de forma a extrair o ano de quitação do imóvel para então ser feita a subtração com relação ao ano identificado na variável Ano Construção Unidade, criando assim a variável “Idade” do imóvel na transação de compra, a qual foi utilizada nos modelos de precificação do imóvel que serão desenvolvidos.

A variável Bairro, utilizada para identificar a localização dos imóveis, apresentou uma ampla variabilidade na base de dados, o que poderia dificultar a análise estatística. Dessa forma, optou-se por realizar uma conversão dessa variável para a respectiva Regional seguindo a divisão territorial vigente na cidade de Belo Horizonte, garantindo uma melhor estruturação dos dados sem perdas significativas de informação.

A conversão foi realizada com base na tabela disponibilizada pela Prefeitura de Belo Horizonte, na qual consta a relação entre os bairros e as regionais do município

(Prefeitura de Belo Horizonte, 2019). A partir dessa correspondência, foi possível agregar os bairros às respectivas regionais, viabilizando uma análise mais consolidada e reduzindo a complexidade do modelo, deixando o mais parcimonioso.

Dado que o objetivo do estudo é avaliar a relação entre as características dos imóveis e sua precificação, e considerando que a variável Regional captura as informações geográficas relevantes, após a sua criação as colunas Endereço, Território e Bairro foram então removidas da base de dados. Essa exclusão visou reduzir o volume de dados processados, garantindo maior eficiência computacional e evitando redundâncias na modelagem preditiva.

2.2 Inclusão de índices macroeconômicos

Além dessa base de dados extraída do portal de dados aberto da prefeitura de Belo Horizonte foram obtidos dados de índices econômicos regionais e nacionais que poderiam compor a análise de precificação dos imóveis, sendo eles o Índice de Preços ao Consumidor Amplo (IPCA) e o componente Habitação deste índice na cidade de Belo Horizonte, compreendendo dados mensais de janeiro de 1994 a dezembro de 2024. Esses dados foram extraídos de arquivos em formatos .pdf publicados no site da Fundação Instituto de Pesquisas Econômicas Administrativas e Contábeis de Minas Gerais, IPEAD.

O Índice Nacional de Preços ao Consumidor (INPC) Brasil compreendendo dados mensais de abril de 1979 a dezembro de 2024, sendo extraído da Tabela 1736, coletada do site do Sistema IBGE de Recuperação Automática, SIDRA.

O índice Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil (SINAPI) relativo ao Estado de Minas Gerais, assim como índice de duas das suas componentes, a Componentes de Materiais de Construção Civil e a Componente de Mão de Obra da Construção Civil, compreendendo dados de janeiro de 1994 a dezembro de 2024, foram extraídos da Tabela 2296, coletada do site do Sistema IBGE de Recuperação Automática, SIDRA.

O índice de inflação do Brasil, compreendendo dados de julho de 1994 a dezembro

de 2024, foi extraído do site do Instituto Brasileiro de Geografia e Estatística, IBGE.

O Índice de Custos da Construção – Capitais (ICCBH), compreendendo dados de julho de 1996 a dezembro de 2024, foi extraído do Instituto Brasileiro de Economia da Fundação Getulio Vargas (FGV-IBRE).

O Índice Geral de Preços (IGP-M) compreendendo dados de janeiro de 1994 a dezembro de 2024, foi extraído do Instituto Brasileiro de Economia da Fundação Getúlio Vargas (FGV-IBRE).

Por fim, o Índice de Variação de Alugueis Residenciais (IVAR) da cidade de Belo Horizonte, compreendendo dados de dezembro de 2018 a dezembro de 2024, foi extraído Instituto Brasileiro de Economia da Fundação Getúlio Vargas (FGV-IBRE).

Todos esses índices foram incorporados à base de dados extraída do portal de dados abertos da prefeitura de Belo Horizonte, a fim de que fossem analisados. Como todos eles são disponibilizados em nível mensal, o valor identificado foi atribuído a todos os imóveis comprados dentro do respectivo mês.

Além disso, com nem todos os índices econômicos estão disponíveis em períodos anteriores a 2018 e, além disso, com o intuito de representar o cenário mais recente do mercado imobiliário em Belo Horizonte, através da Data Quitação Transição, foi realizado um filtro para imóveis quitados apenas depois de 2019. Dessa forma, a modelagem preditiva realizada compreende identificar as movimentações de preço entre os anos de 2019 e 2024. Neste intervalo de tempo de 6 anos, o total de transações foi de 67.376, representando o contexto mais recente da base para identificar de forma mais próxima a realidade de preços praticados no mercado imobiliário na cidade de Belo Horizonte.

2.3 Estruturação dos dados

Além dos filtros nos tipos de imóveis e abrangência temporal, devido a problemas gerados por valores atípicos (outliers) identificados numa primeira análise, optou-se por aplicar um filtro na variável Valor Declarado de forma a restringir as análises apenas para imóveis residenciais comprados na faixa de preço entre R$100.000,00 (cem mil reais) e R$1.500.000,00 (um milhão e meio de reais).

Por fim, foi aplicado um filtro com base na razão entre os preços dos imóveis vendidos, indicados pela variável Valor Declarado, e o valor venal do imóvel estimulado pela prefeitura na cobrança do ITBI presente na variável Valor Base Cálculo. Considerou-se apenas imóveis para os quais o valor desta razão ficou acima de 0,5, ou seja, o valor de venda foi, no mínimo, 50% do valor venal do imóvel. O objetivo deste filtro é mitigar ações de transações de imóveis que fogem da compra e venda tradicional de mercado, ações estas como doações de imóveis com registro de preço muito abaixo do real, compras de leilões com preços muito abaixo do valor de mercado do imóvel, e compra de imóveis na planta cujo valor é bem abaixo do valor de venda após estar construído, sendo a precificação do valor venal pela prefeitura o projeto final do imóvel após construído.

Assim, após a aplicação de todos os filtros/critérios de exclusão mencionados acima, a base de dados final utilizada contou com as transações de venda de imóveis residenciais realizadas entre janeiro de 2019 e maio de 2024 no município de Belo Horizonte, com valor de venda entre R$100.000,00 e R$1.500.000,00, fabricados a partir de 1994 e cujo valor de venda não fosse inferior a 50% do valor venal declarado para o imóvel.

Para o ajuste dos modelos, a base de dados foi dividida por meio da função initial_split() pertencente ao pacote rsample, em duas amostras: uma para treinamento (70% dos dados) e outra para teste (30%). A fim de garantir um equilíbrio adequado entre os conjuntos, foi aplicada a técnica de estratificação utilizando a função interaction() pertencente ao pacote básico da linguagem R concatenando as variáveis categóricas: Regional, Tipo Construtivo, Padrão de Acabamento e Faixa Etária do Imóvel. Esta última foi definida de modo que contemplasse imóveis menores que 10 anos, entre 10 e 20 anos, entre 20 e 30, e também entre 30 e 40 anos, gerando assim apenas 4 tipos de faixas etárias que foram combinadas com as demais variáveis e facilitaram a concatenação, ao invés de utilizar as várias idades diferentes dos imóveis presentes na base. Esta concatenação das variáveis categóricas foi utilizada na função *mutate()* do pacote *dplyr* do *software* R.

O uso desse procedimento durante a separação dos dados assegura que ambas as amostras, treino e teste, mantenham a proporção original das diferentes categorias,

reduzindo possíveis vieses na modelagem. Além disso, a estratificação permite que cada combinação única dessas variáveis categóricas seja tratada como uma unidade distinta, garantindo que sua distribuição seja preservada nos conjuntos de treino e teste. Dessa forma, a divisão dos dados foi conduzida de maneira estruturada e estatisticamente adequada, assegurando a representatividade das observações em ambos os subconjuntos.

2.4 Análise Exploratória dos Dados

A análise exploratória de dados constitui uma etapa inicial fundamental e preliminar no processo de modelagem preditiva, compreendendo a estrutura, os padrões e características subjacentes dos dados antes de aplicar técnicas estatísticas mais sofisticadas. Como enfatizado por Turkey (1997), a análise exploratória prioriza a descoberta de insights e hipóteses a partir dos dados, utilizando uma abordagem sistemática e visual. Uma análise exploratória eficiente segue uma estrutura que inclui algumas etapas descritas a seguir.

Para a variável dependente, preço dos imóveis, segundo Hair (2019), as estatísticas descritivas básicas fornecem um sumário numérico inicial que orienta análises subsequentes. O processo de análise descritiva dos dados, inicia-se com o cálculo de estatísticas descritivas básicas na variável equivalente a ela na base de dados. Estatísticas como as medidas de tendência central sendo média e mediana, de dispersão como o desvio-padrão, variância e amplitude interquartil além de medidas da forma de distribuição como assimetria e curtose.

As avaliações numéricas aliadas a histogramas compostos com gráficos de densidade permitem examinar a forma da distribuição, identificando características como assimetria e presença de valores extremos. Os gráficos de boxplots também complementam a análise, pois oferecem representação visual de estatísticas como a mediana e os quartis. As visualizações da distribuição de preço de compra dos imóveis podem revelar concentrações em determinadas faixas de valores ou diferenças distributivas entre as características regionais e construtivas destes imóveis.

Além das análises gráficas e resumos numéricos univariados das variáveis

envolvidas no estudo, a exploração de relações entre a variável resposta e as variáveis independentes, bem como das variáveis explicativas entre si, compõe um outro estágio da análise descritiva dos dados. Para as variáveis numéricas da base de dados, a matriz de correlação oferece uma visão geral da força e a direção da relação linear entre as variáveis, transformada em um gráfico complementada por correlograma e seu *heatmap* de correlações, facilita a identificação de colinearidade entre variáveis.

No âmbito de um modelo de regressão, a multicolinearidade entre variáveis preditoras merece atenção especial. De acordo com KUTNER (2013), correlações elevadas entre variáveis podem comprometer a estabilidade e interpretabilidade dos modelos de regressão. Podendo ser identificada no gráfico de correlações com dados obtidos da matriz de correlações, e através do Fator de Inflação da Variância, VIF, que surge como métrica formal para quantificar este fenômeno, denominado por:

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

onde R_j^2 representa o coeficiente de determinação quando uma variável preditora x_j é regredida contra todas as demais variáveis independentes (preditoras) no modelo. Os valores de VIF superiores a 10 geralmente indicam multicolinearidade problemática, enquanto valores acima de 5 já merecem atenção na sua utilização.

2.5 Modelo de Regressão Linear Múltipla

O modelo de regressão linear múltipla (RLM) permite a incorporação de múltiplas variáveis independentes a fim de explicar a variabilidade de uma variável dependente. Este modelo, conforme apresentado em Montgomery (2021), é expresso como:

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \cdots + \beta_P.X_P + \varepsilon$$

onde Y representa a variável dependente, X_1, X_2, ..., X_p são as variáveis independentes, β_0, β_1, β_2, ..., β_p são os parâmetros do modelo, e ε é o termo de erro aleatório.

As suposições de que a variabilidade dos erros é constante para i=1, ..., n, de que eles são independentes e que o erro segue uma distribuição normal são relevantes para validar os testes de significância e intervalos de confiança que podem ser conduzidos ao longo da análise (MONTGOMERY et al., 2021).

Baseado nos resíduos do modelo, definidos através da diferença entre valores reais e preditos, há métodos descritivos e inferenciais voltados para avaliação das suposições feitas para o termo de erro. Entretando, como o interesse principal da análise está voltado na determinação de predições, é comum que a avaliação de tais suposições seja de certo modo negligenciado em detrimento à avaliação de métricas do modelo preditivo, como considerado neste estudo.

Para a análise e definição de modelos preditivos, neste trabalho, todas as análises predições que envolvem modelo de regressão foram feitas usando o *software* R (R CORE TEAM, 2024).

2.6 Random Forest

O algoritmo Random Forest, introduzido por Breiman (2001), representa uma técnica de *ensemble* baseada na agregação de múltiplas árvores de decisão, combinando os princípios de *bagging* (*bootstrap aggregating*) com a seleção aleatória de variáveis. Buscando assim superar as limitações de métodos preditivos baseados em árvores individuais.

Na modelagem de relações da variável resposta o Random Forest pode identificar a contribuição relativa de diferentes fatores, como variáveis quantitativas, categóricas a respeitos dos imóveis assim como indicadores macroeconômicas do período de compra do imóvel. Fornecendo uma hierarquia de importância baseada em critérios objetivos de redução de erro, permitindo esta hierarquização, focar a atenção nas variáveis mais relevantes, compreendendo melhor a estrutura subjacente da base em estudo.

Para a análise e definição de modelos preditivos, neste trabalho, todas as análises

predições que envolvem modelo de Random Forest foram feitas utilizando o pacote RandomForest (LIAW E WIENER, 2023) disponível no *software* R.

2.7 Métricas de Avaliação para Modelos Preditivos

A avaliação da capacidade preditiva de modelos é um passo importante na elaboração de sistemas preditores confiáveis. Dentre os indicadores de avaliação, tem-se o Erro Quadrático Médio (MSE), a Raiz Quadrada do Erro Quadrático Médio (RMSE), o Erro Absoluto Médio (MAE), o Erro Percentual Absoluto Médio (MAPE), o coeficiente de determinação (R^2) e o coeficiente de determinação ajustado (R^2 ajustado).

O Erro Quadrático Médio, MSE representa uma medida que penaliza os erros maiores, já o RMSE os maiores erros, desproporcionalmente, devido a sua raiz quadrada aplicada à função quadrática, sendo definido como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Nesta equação, y_i representa o valor observado e $\hat{y}_i$ o valor predito. Em análises do valor declarado de compra do imóvel, o RMSE indica a magnitude média dos erros de predição, com valores menores representando maior precisão do modelo. Segundo Willmott e Matsuura (2005), embora o RMSE seja matematicamente tratável e tenha interpretação clara em termos da variância residual, sua sensibilidade a outliers pode distorcer a avaliação geral do modelo em certas situações.

O Erro Absoluto Médio, por tratar todos os erros proporcionalmente apresenta-se como uma alternativa robusta na avaliação da qualidade do modelo, sendo definido como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Este indicador representa a média das diferenças absolutas entre valores

observados e preditos, fornecendo uma medida do tamanho do erro. Como ele não eleva os desvios ao quadrado, o MAE é menos sensível a valores extremos quando comparado ao RMSE, oferecendo uma percepção equilibrada do desempenho do modelo preditivo quando a distribuição dos erros apresenta outliers.

O Erro Percentual Absoluto Médio, MAPE, é um indicador capaz de expressar percentualmente o erro da predição, sendo definido por:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Para o valor de compra de um imóvel, o MAPE é indica o percentual médio de erro nas predições, facilitando a interpretação dos erros em termos percentuais e permitindo comparações entre modelos aplicados a diferentes escalas ou variáveis dependentes. Entretanto, como observado por Hyndman e Koehler (2006), o MAPE pode apresentar problemas quando os valores observados se aproximam de zero ou apresentam grande variabilidade na escala.

O coeficiente de determinação, R^2, é um indicador que quantifica a proporção da variabilidade na variável resposta explicada pelo modelo, sendo definida por:

$$R^2 = 1 - \frac{SSres}{SStot} = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \hat{y})^2}$$

Em que, SSres é a soma dos quadrados dos resíduos e SStot a soma total dos quadrados, com resultados variando entre 0 e 1, com valores mais próximos de 1 indicando maior capacidade explicativa. Este indicador oferece uma capacidade explicativa do modelo, sendo particularmente útil para comunicar resultados a públicos diversos. No entanto, James et al. (2017) alerta que o R^2 calculado na amostra de treinamento pode superestimar o desempenho real do modelo devido ao sobre ajuste.

Além do R^2, o indicador R^2 ajustado (R^2adj) representa um indicador que penaliza a complexidade do modelo, sendo apresentado como:

$$R^2 Ajust = 1 - (1 - R^2).\frac{(n-1)}{(n-p-1)}$$

Este indicador permite evitar a seleção de modelos excessivamente complexos baseados apenas na estimação do R² convencional. Kuhn e Johnson (2018) chamam a atenção para a importância de avaliar estas métricas em dados não utilizados no treinamento dos modelos para obter estimativas realistas da capacidade preditiva.

Em resumo o uso de métricas adequadas na avaliação do modelo preditivo, devem considerar as características específicas da análise em questão e os objetivos da modelagem. Enquanto o RMSE e o MAE fornecem perspectivas complementares sobre o tamanho absoluto dos erros, o MAPE oferece uma visão relativa sobre eles, e por fim, o R² é capaz de quantificar a capacidade explicativa global do modelo. A análise conjunta destas métricas proporciona uma avaliação robusta e adequada sobre a qualidade preditiva dos modelos.

3 RESULTADOS

3.1 Analise Descritiva variável resposta

Apos os tratamentos de dados na base, iniciou-se a análise exploratória da variável resposta (Valor Declarado). A análise estatística do valor declarado de venda dos imóveis conta com 67.376 observações. Conforme mostrado na Tabela5, a média (R$431.085) e mediana (R$350.000) apresentam valores razoavelmente próximos, evidenciando uma distribuição assimétrica a direita (Figura 1), não seguindo uma distribuição normal. Entretanto, a avaliação mais detalhada dos parâmetros de forma evidencia uma assimetria positiva (1,45), indicando uma concentração de observações nos valores inferiores da distribuição, com uma cauda mais alongada em direção aos valores maiores.

Figura 1: Histograma da variável resposta (Valor Declarado)

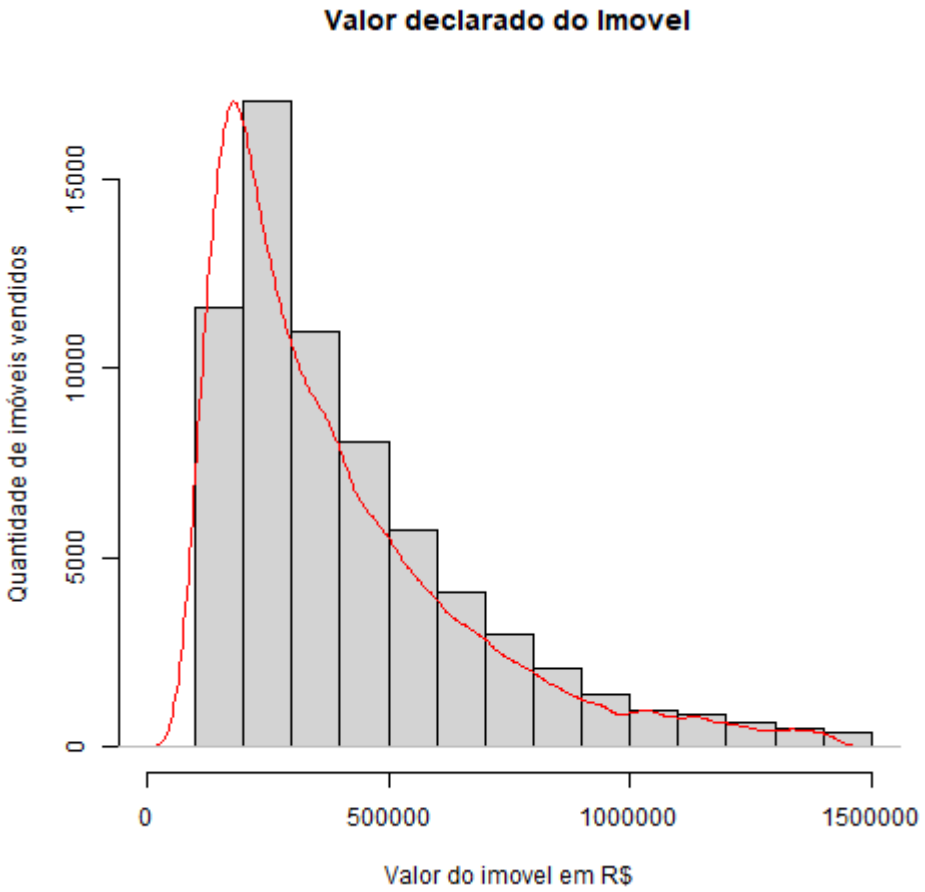


Tabela 5: Estatísticas descritivas da variável resposta.

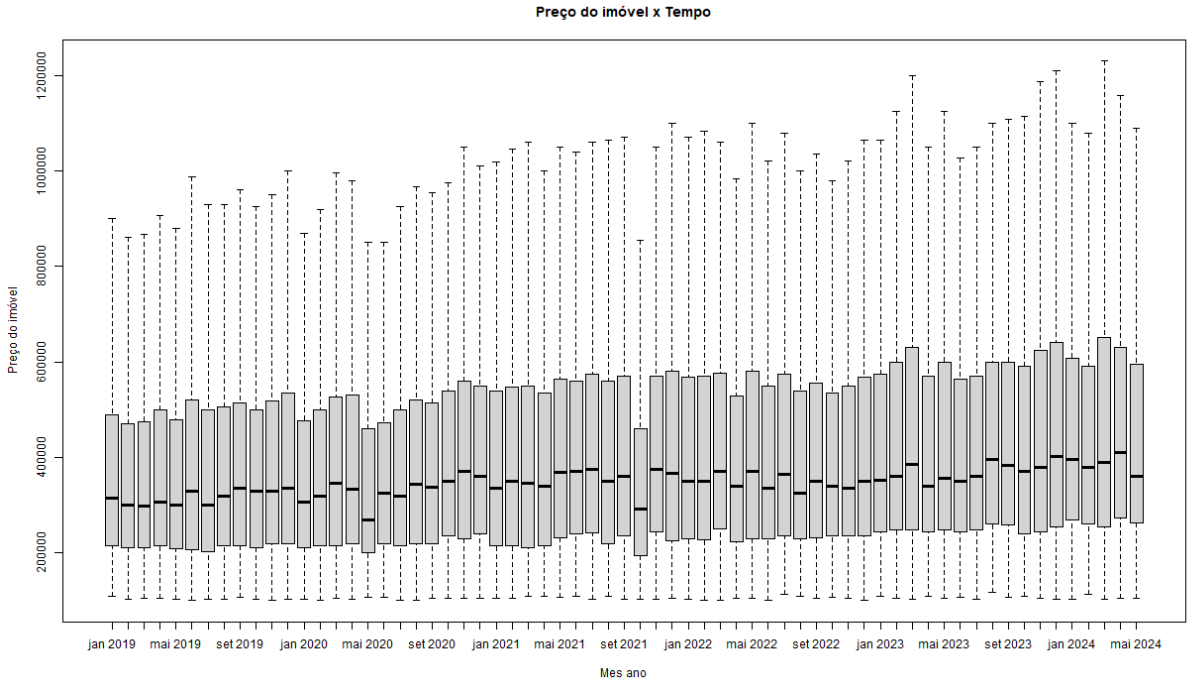
Valor Declarado	
Número de observações	67.376
Média	431.085
Mediana	350.000
Desvio padrão	272.147
Mínimo	100.310
Máximo	1.499.900
Primeiro quartil (Q1)	230.000
Terceiro quartil (Q3)	550.000
Amplitude interquartil (IQR)	320.000
Assimetria	1,45
Curtose	1,91

A distribuição dos valores declarados apresenta uma considerável dispersão, conforme indicado pelo desvio-padrão de R$272.147 e pela amplitude interquartil de R$320.000. Esses valores evidenciam uma significativa variabilidade nos dados, sugerindo a presença de uma ampla gama de observações. O coeficiente de curtose, igual a 1,91 caracteriza a distribuição como mesocúrtica, ou seja, sem uma concentração excessiva de valores em torno da média.

A assimetria positiva (1,45) indica que a distribuição é enviesada à direita, o que sugere que há uma maior frequência de valores inferiores à média e a presença de observações elevadas que prolongam a cauda direita da distribuição. Essa característica também é refletida na diferença substancial entre a média (R$431.085) e a mediana (R$350.000), evidenciando o impacto de valores extremos na medida de tendência central. Assim, a análise descritiva sugere que a variável "Valor Declarado" não segue uma distribuição perfeitamente simétrica, sendo influenciada por valores altos que deslocam a média para cima.

A análise da temporal da variável resposta, é dada pelo boxplot apresentado na Figura 2. Os resultados indicam uma tendência de valorização dos imóveis ao longo do período. A mediana do valor declarado passou de R$315.000,00 em 2019 para R$359.500,00 em 2024, enquanto a média cresceu de R$392.274,50 para R$ 463.743,30. O desvio-padrão também aumentou, evidenciando maior heterogeneidade nos preços, possivelmente devido à diversificação do mercado. Essa análise demonstra que o mercado imobiliário apresentou valorização contínua, com aumento na dispersão dos valores, indicando um ambiente mais segmentado. Esses dados são essenciais para orientar decisões de investidores e profissionais do setor.

Figura 2: Boxplot do Valor Declarado no tempo.



A relação da variável resposta Valor Declarado com a variável Regional, é apresentada na Figura 3 e algumas medias descritivas separadas por região são apresentadas na Tabela 6. A região Centro-Sul se destaca não apenas pelos maiores valores medianos (R$740.000) e médios (R$770.937,2), mas também pelo maior desvio padrão (R$322.069,2), indicando uma valorização imobiliária superior e uma grande variação nos preços dos imóveis. Esse comportamento pode ser atribuído à infraestrutura consolidada, maior demanda e localização privilegiada, fatores que favorecem a apreciação dos imóveis na região.

Em contrapartida, a região Venda Nova apresenta os menores valores mediano (R$ 215.725,5) e médio (R$245.759,6), o que sugere uma valorização inferior dos imóveis. Esse cenário pode ser explicado pela localização periférica e pela menor concentração de imóveis de alto padrão. A menor valorização nessa área reflete as características específicas do mercado local, que pode ter uma oferta mais voltada para imóveis de menor valor.

Em diversas regionais, observa-se uma proximidade significativa entre os valores médios e medianos, o que sugere distribuições de preços tendendo à simetria. Esse padrão é característico de mercados imobiliários relativamente previsíveis, onde as distorções de preços dentro de cada região são limitadas, proporcionando um cenário mais estável para compradores e investidores.

As regiões Pampulha, Oeste e Leste apresentam medianas semelhantes, variando entre R$345.000 e R$400.000. Esse comportamento sugere que essas áreas possuem padrões de valorização imobiliária relativamente alinhados, o que pode indicar uma similaridade nos perfis de compradores e nas estruturas de mercado dessas regiões. A uniformidade nos valores medianos também pode indicar uma estabilidade no mercado local, que apresenta características semelhantes em termos de oferta e demanda.

Os dados analisados indicam que a região Centro-Sul se destaca em termos de valorização imobiliária, com valores medianos e médios mais elevados. Por outro lado, a região Venda Nova apresenta os menores valores, refletindo um mercado com menor valorização. A semelhança entre as médias e medianas nas diversas regionais reforça a previsibilidade do mercado.

Figura 3: Boxplot do Valor Declarado por Regional.

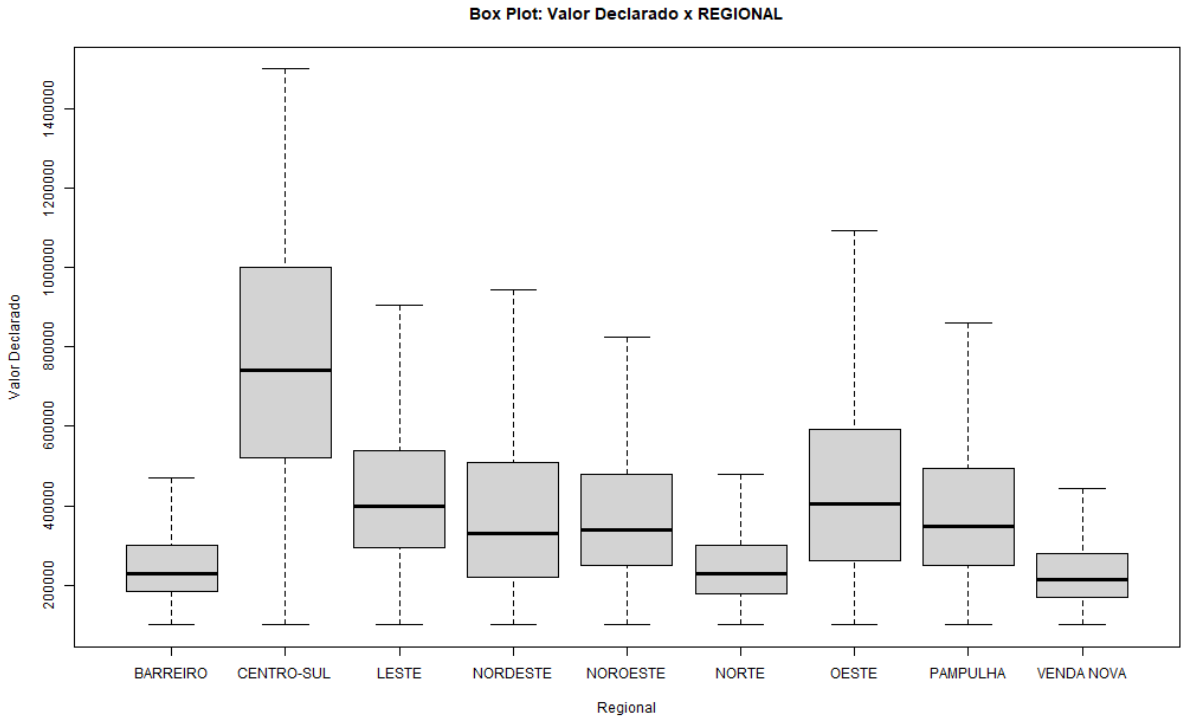
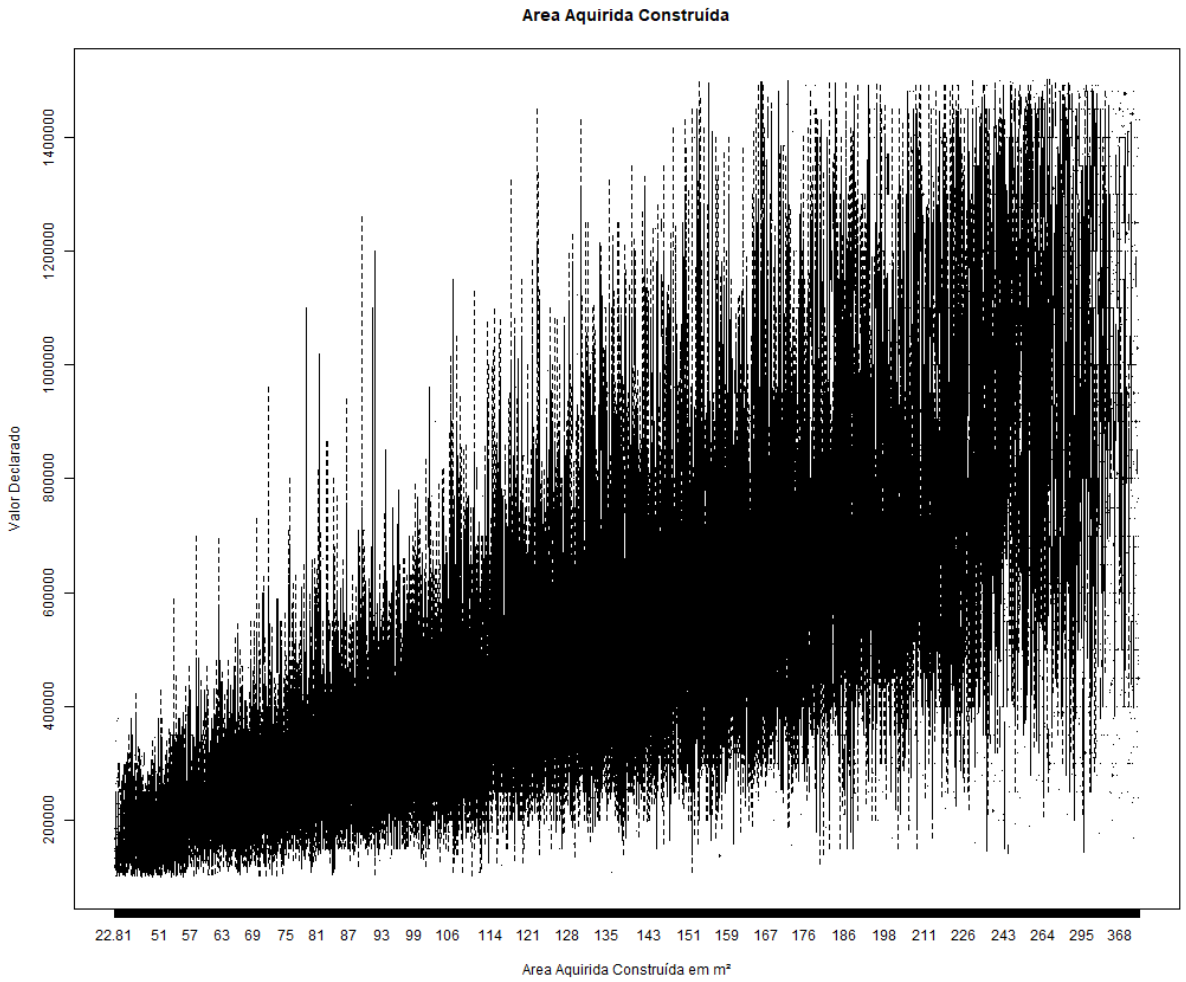


Tabela 6: Estatísticas descritivas do Valor Declarado por Regional.

Estatísticas Descritivas do Valor Declarado por Regional							
Regional	**Mínimo**	**Q1**	**Mediana**	**Média**	**Q3**	**Máximo**	**Desv Pad**
BARREIRO	101.074	190.000	233.273	268.191,4	300.00	1.440.000	133.186,7
CENTRO-SUL	101.825	525.000	740.000	770.937,2	1.000.000	1.499.900	322.069,2
LESTE	102.000	300.000	400.000	435.903,5	540.000	1.475.000	187.351,3
NORDESTE	101.000	215.000	310.000	394.264,4	500.000	1.491.000	248.303,2
NOROESTE	102.250,6	250.000	340.000	391.418,5	480.000	1.485.000	201.022,9
NORTE	102.064,5	180.000	222.669,7	257.770,6	300.000	1.300.000	119.255
OESTE	101.450,6	258.500	400.000	458.127,7	588.000	1.495.000	255.291,3
PAMPULHA	101.472,9	248.000	345.000	400.894,6	495.000	1.499.000	214.890,6
VENDA NOVA	100.310,2	178.000	215.725,5	245.759,6	280.000	1.300.000	108.675

A relação da variável resposta Valor Declarado com a variável Área Construída Adquirida, é apresentada na Figura 4. Através do boxplot gerado, é perceptível a variabilidade de dados relativos a área construída adquiridas, da base de dados. É possível perceber uma relação crescente entre a área construída adquirida e o preço pelo qual o imóvel foi vendido. Imoveis com áreas construídas menores, possuem preços menores, e imóveis maiores, com mais áreas construídas tem valores mais elevados, indicando uma correlação proporcional entre essas variáveis.

Figura 4: Boxplot do Valor Declarado por Área Construída Adquirida.



A relação da variável resposta Valor Declarado com a variável Tipo Construtivo, é apresentada na Figura 5 e na Tabela 7. A análise da relação entre o Valor Declarado e os dois tipos construtivos considerados no estudo, apartamento e casa, revela pequenas diferenças nas distribuições de preços entre eles, refletindo as características específicas de cada tipo de imóvel.

Para os apartamentos, a mediana é de R$340.000, com uma média de R$ 428.441,2, indicando que a maioria dos apartamentos está concentrada em valores abaixo da média, com alguns imóveis de valor mais alto elevando a média. O desvio padrão de R$273.792,9 sugere uma variação significativa nos preços, indicando a presença de imóveis tanto mais acessíveis quanto de alto padrão dentro da categoria de apartamentos. A amplitude dos valores é considerável, com o valor máximo atingindo R$1.499.900, refletindo imóveis de grande valor.

Por outro lado, os casas apresentam uma mediana de R$350.000, ligeiramente superior à dos apartamentos, e uma média de R$425.636,3, também mais baixa do que o valor máximo, indicando uma distribuição similar com alguns imóveis de valor elevado. O desvio padrão de R$253.489,4 também é considerável, embora ligeiramente inferior ao dos apartamentos, sugerindo uma variação de preços um pouco mais controlada dentro dessa categoria. O valor máximo registrado para as casas foi de R$1.484.900, o que ainda é elevado, mas não ultrapassa o valor máximo dos apartamentos.

Esses dados indicam que, embora ambas as categorias de imóveis apresentem uma grande dispersão de preços, os apartamentos possuem uma variabilidade um pouco maior e uma concentração de valores mais altos, enquanto as casas mostram uma distribuição de preços ligeiramente mais homogênea, com menores variações em relação ao valor médio. Esses padrões podem ser úteis para entender as preferências e a dinâmica de mercado orientando investidores e compradores sobre as possibilidades de valorização em cada segmento.

Figura 5: Boxplot do Valor Declarado pelo Tipo Construtivo do imóvel.

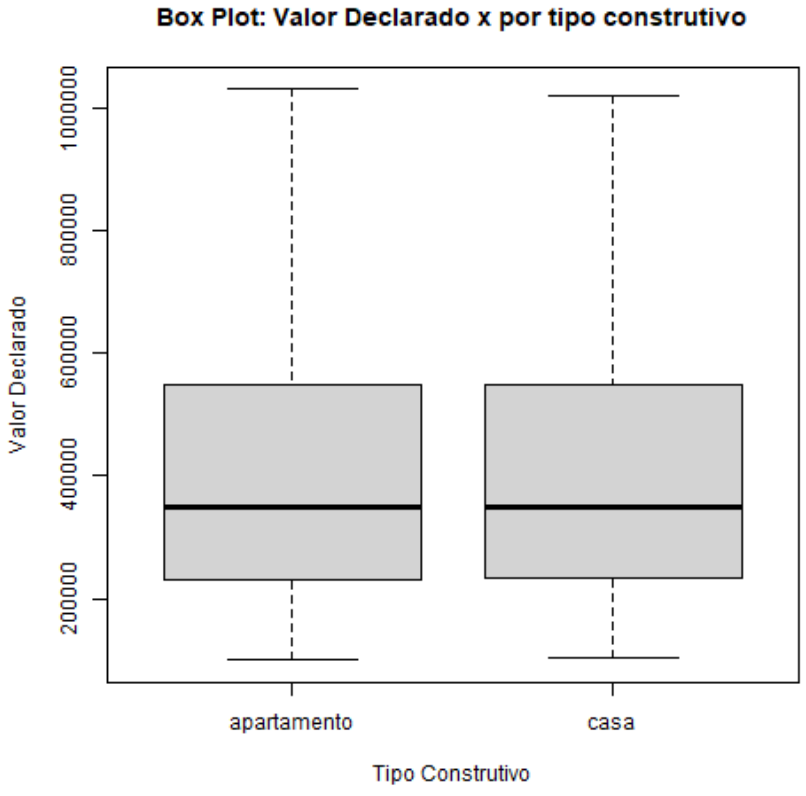


Tabela 7: Estatísticas descritivas do Valor Declarado pelo Tipo Construtivo do imóvel.

Estatísticas Descritivas do Valor Declarado pelo Tipo Construtivo do imóvel

Tipo Construtivo	Mínimo	Q1	Mediana	Média	Q3	Máximo	Desv Pad
APARTAMENTO	100.310,2	227.424,6	340.000	428.441,2	550.000	1.499.900	273.792,9
CASA	102.323	240.000	350.000	425.636,3	550.000	1.484.900	253.489,4

A Figura 6 e a Tabela 8 apresentam a relação da variável resposta Valor Declarado com a variável Padrão Acabamento. A análise revela uma relação clara entre a qualidade construtiva e a valorização imobiliária. Imóveis com padrão Muito Simples apresentam os menores valores medianos (R$150.000) e médios (R$198.303,6), refletindo um menor nível de investimento e em alguns casos, uma localização menos valorizada. Já os imóveis de padrão Muito Bom exibem os maiores valores medianos (R$846.500) e médios (R$855.974,5), evidenciando uma maior valorização, possivelmente atrelada a melhores infraestruturas e localização privilegiada.

A variabilidade dos valores acompanham o padrão de acabamento, conforme indicado pelos desvios padrão a medida em que se melhora o padrão de acabamento dos imóveis. Aqueles com padrão Muito Simples possuem menor variação (R$131.604,4), enquanto aqueles classificados com padrão Muito Bom apresentam um desvio padrão elevado (R$379.681,8), sugerindo uma maior heterogeneidade de preços dentro dessa categoria. Esse comportamento pode estar relacionado à diversidade de características e localizações dos imóveis de alto padrão.

Dessa forma, os dados evidenciam a influência do padrão de acabamento na precificação dos imóveis, destacando que construções de maior qualidade tendem a apresentar maior valorização e variabilidade nos preços, enquanto imóveis de padrões inferiores possuem uma precificação mais homogênea e limitada.

Figura 6: Boxplot do Valor Declarado pelo Tipo de Padrão de Acabamento do imóvel.

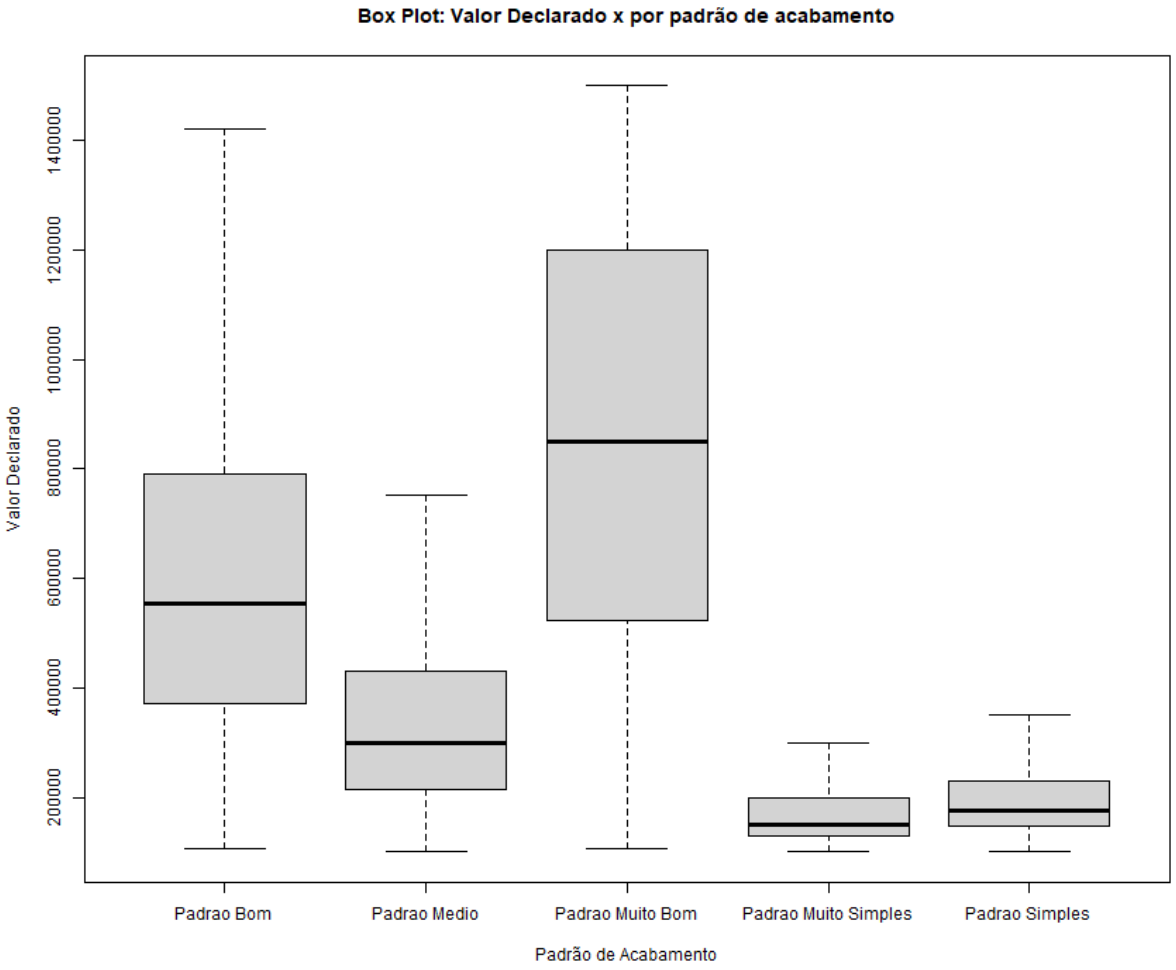


Tabela 8: Estatísticas descritivas do Valor Declarado pelo Tipo de Padrão de Acabamento do imóvel.

Estatísticas Descritivas do Valor Declarado pelo Padrão de Acabamento do imóvel							
Padrão de Acabamento	**Mínimo**	**Q1**	**Mediana**	**Média**	**Q3**	**Máximo**	**Desv Pad**
MUITO SIMPLES	100.664,4	130.000	150.000	198.303,6	200.000	1.100.000	131.604,4
SIMPLES	101.000	150.000	175.000	213.959,1	230.000	1.444.500	125.295,8
MÉDIO	100.310,2	215.000	290.000	342.639,1	420.000	1.491.000	180.314,8
BOM	107.323,6	350.000	545.000	595.407,6	780.000	1.499.092	300.250
MUITO BOM	105.960	521.185	846.500	855.974,5	1.200.000	1.499.900	379.681,8

A analise a variável resposta Valor Declarado com a variável Idade Imóvel, é apresentada na Figura 7. A relação entre a idade do imóvel e seu valor declarado apresenta um comportamento não linear, sugerindo que diferentes faixas etárias podem influenciar a precificação de maneira distinta. Pode-se observar que imóveis mais novos entre 1 e10 anos, apresentam valores medianos e médios crescentes, atingindo um pico por volta dos 10 anos, onde a mediana alcança R$400.000,00 e a

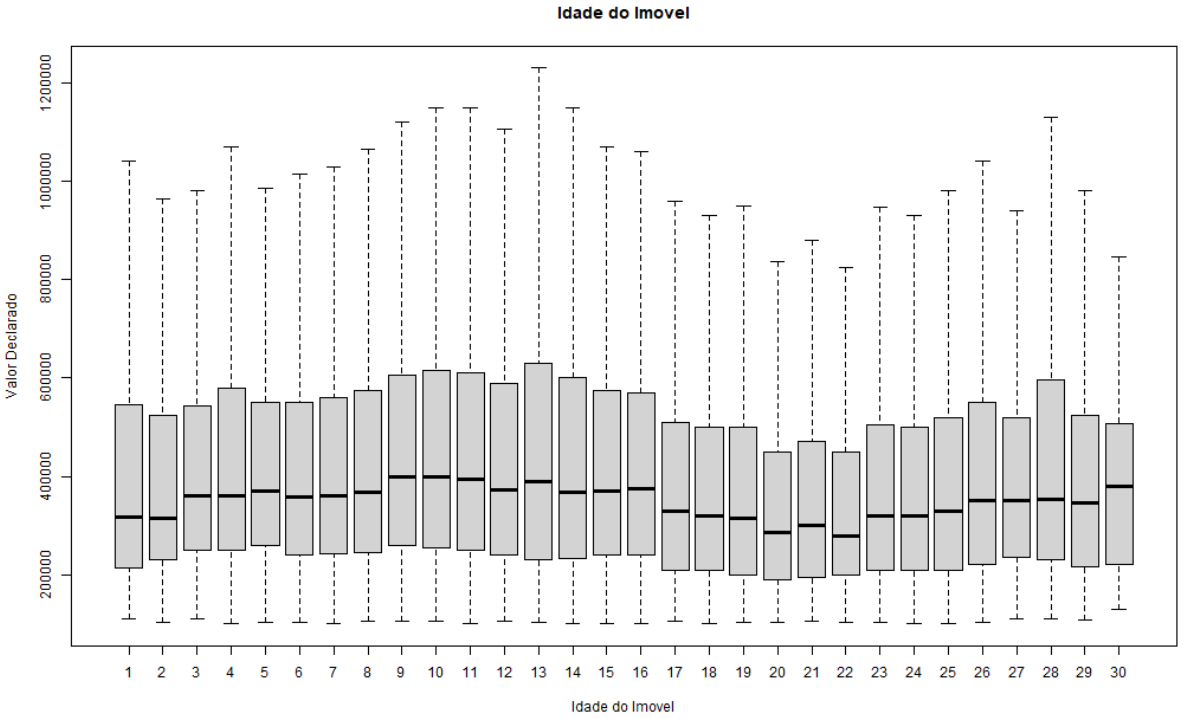
média em torno de R$471.558,60. Esse comportamento pode estar relacionado à valorização inicial de empreendimentos novos, impulsionada por fatores como modernidade, menor necessidade de manutenção e infraestrutura atualizada.

A partir dos 11 anos, os valores começam a oscilar, mas sem uma tendência clara de depreciação acentuada. Imóveis entre 11 e 20 anos mantêm valores medianos relativamente estáveis, variando entre R$315.000,00 e R$600.000,00, sugerindo que a qualidade construtiva e possíveis reformas podem mitigar a perda de valor ao longo do tempo.

Já para imóveis acima de 20 anos, observa-se uma leve redução na mediana e na média, indicando que a idade pode impactar na precificação, principalmente em construções mais antigas que não passaram por modernizações. Ao observar a dispersão dos valores nessa faixa etária, percebe-se uma variação em torno de R$ 246.133,70 a quase R$600.000, evidenciando um comportamento com alta heterogeneidade nos valores de mercado dentro de cada faixa etária dos imóveis, possivelmente refletindo diferenças na localização, estado de conservação e características específicas deles.

Em resumo, há um indicativo de imóveis novos apresentarem maior valorização inicial, enquanto a estabilização dos valores ao longo do tempo sugere que fatores como manutenção e localização são mais determinantes do que a idade isoladamente.

Figura 7: Boxplot do Valor Declarado pela idade do imóvel.

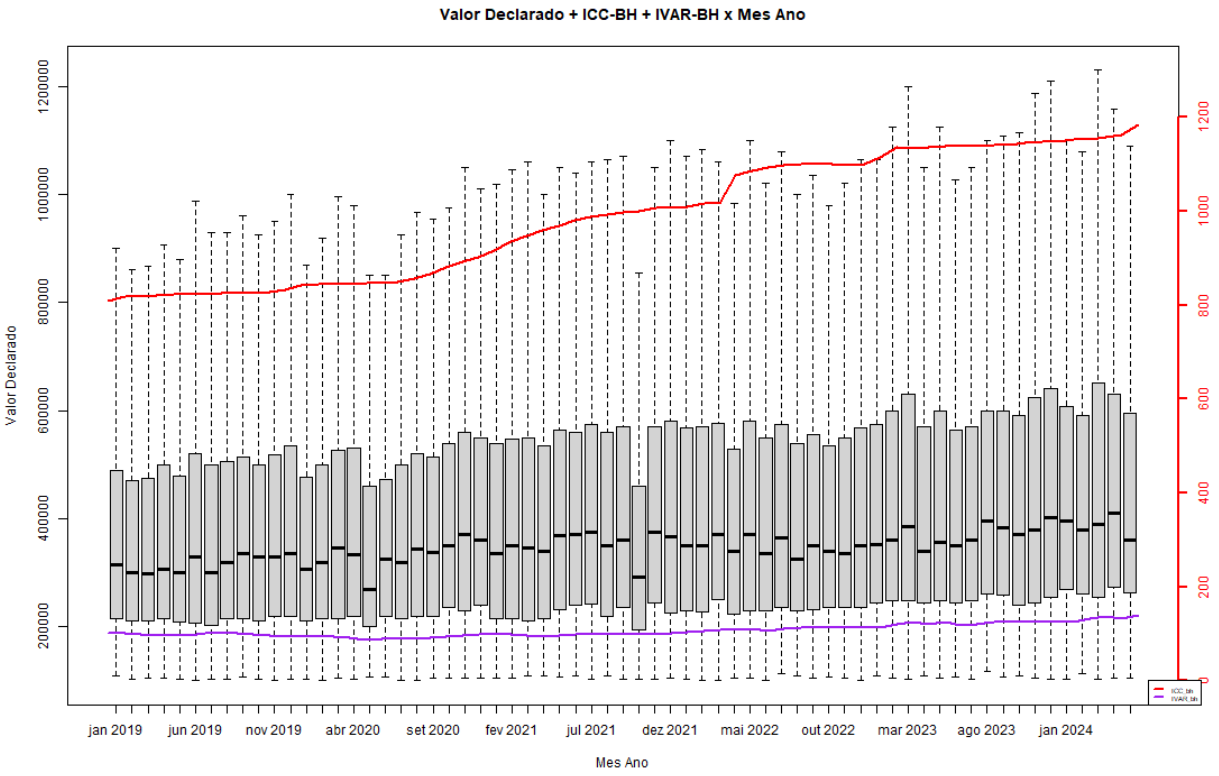


Analisando a variável resposta Valor Declarado com relação aos indicadores ICC e IVAR. Através do boxplot exibido na Figura 8 juntamente com os índices ICC e IVAR ao longo do tempo, percebe-se uma variação de preço do imóvel, discreta mês a mês. Sendo a média de preços levemente acompanhada do índice de construção civil até março 2022, e a partir de abril de 2022, o índice de construção civil sobre um aumento cujo preço do imóvel não acompanha o mesmo ritmo graficamente.

A partir de abril de 2025, graficamente percebe-se um levíssimo aumento no índice de variação de alugueis, ligeiramente acompanhado pelo preço dos imóveis na cidade de Belo Horizonte.

Estes aumentos a partir de abril de 2025, pós-pandemia da Covid-19, pode indicar um possível aumento no preço de venda dos imóveis devido a uma alta no preço dos aluguéis acompanhado também pelo aumento de custos na construção civil, como mostrado graficamente na linha correspondente ao ICC no gráfico.

Figura 8: Boxplot do Valor Declarado pelos índices ICC e IVAR.



A correlação entre variáveis é um dos principais métodos estatísticos para identificar relações lineares entre fatores que podem influenciar o preço dos imóveis, além de permitir análise de possível multicolinearidade entre as variáveis preditoras. Através dos resultados exibidos na Figura 9 é possível perceber que a variável Área Construída Adquirida possui uma correlação positiva forte com a variável Valor Declarado (0.77), sugerindo que o tamanho da área construída adquirida tem impacto direto no valor do imóvel.

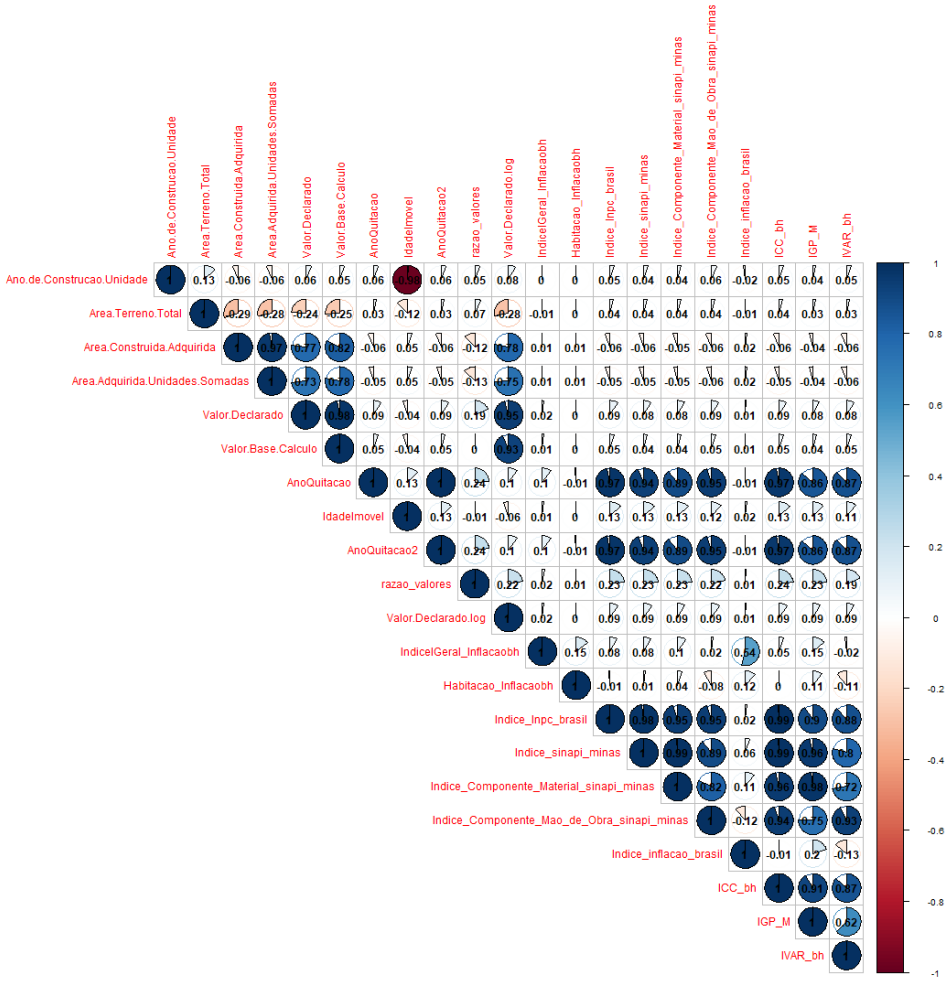
Avaliando os indicadores macroeconómicos, como os índices, percebe-se que o índice ICC, referente aos custos da construção civil demonstram uma leve relevância na precificação imobiliária (0.09). Já os índices de Componente Material Sinapi minas e Componente de Mão de Obra Sinapi Minas, apresentam leves correlações positivas a variável resposta (0.08 e 0.09 respectivamente). E o índice de variação de alugueis residenciais, possui uma correlação positiva leve (0.08) com o preço do imóvel, evidenciando que variações nos custos dos materiais, da mão de obra, e o preço dos aluguéis podem impactar levemente o preço dos imóveis no mercado imobiliário de forma direta.

Por outro lado, a variável Idade Imóvel apresenta uma correlação levemente

negativa com a variável Valor Declarado (-0.042), indicando que imóveis mais antigos por apresentarem depreciação ou obsolescência das construções ao longo do tempo, podem influenciar negativamente no preço do imóvel.

De maneira geral, os resultados apontam que variáveis relacionadas à área dos imóveis podem influenciar mais sobre o preço do imóvel, enquanto fatores macroeconômicos e a idade do imóvel apresentam uma correlação mais sutil, tendo uma leve interferência no preço dos imóveis.

Figura 9: *Correlation plot* das correlações entre as variáveis quantitativas.



3.2 Ajuste dos modelos preditivos

Usando o conjunto de dados de treino (70% dos dados), o desenvolvimento de modelos preditivos abordou, inicialmente, a metodologia de empregar um modelo de regressão linear da variável resposta com todas as variáveis dependentes categóricas e com as variáveis quantitativas com mais relevância na análise de

correlações. Aplicando um modelo linear de forma separada a cada uma delas, a fim de avaliar o comportamento individual com a variável resposta, foram obtidos os resultados exibidos na Tabela 9.

Tabela 9: Estatísticas do modelo linear aplicado separadamente as variáveis independentes.

Estatísticas do modelo linear aplicado separadamente as variáveis independentes										
Variável	**R^2**	**R^2 ajustado**	**Erro Padrão da Estimativa**	**Estatística F do modelo**	**P-valor**	**Graus de Liberdade**	**Log-Likelihood**	**AIC**	**BIC**	**Desvio-padrão do erro**
ÁREA CONSTRUIDA ADQUIRIDA	0,6040	0,6040	170.700,2	71.947,95	0	1	-635124	1270255	1270281	$1,37x10^{15}$
ICC_BH	0,0079	0,0079	270.197,6	377,95	7,39E-84	1	-656784	1313573	1313600	$3,44\ x10^{15}$
IDADEIMÓVEL	0,0019	0,0019	271.013,7	92,09	8,64E-22	1	-656926	1313858	1313884	$3,46\ x10^{15}$
COMPONENTE DE MAO DE OBRA (SINAPI MINAS)	0,0072	0,0072	270.292,8	344,46	1,28E-76	1	-656800	1313607	1313633	$3,45\ x10^{15}$
COMPONENTE DE MATERIAL (SINAPI MINAS)	0,0067	0,0067	270.354,4	322,83	6,09E-72	1	-656811	1313628	1313654	$3,45\ x10^{15}$
INFLACAO-BH	0,0005	0,0005	271.206	25,09	5,47E-07	1	-656959	1313925	1313951	$3,48\ x10^{15}$
IVAR-BH	0,0065	0,0065	270.383,2	312,70	9,43E-70	1	-656816	1313638	1313665	$3,46\ x10^{15}$
PADRÃO DE ACABAMENTO	0,3013	0,3013	226.748,9	5086,21	0	4	-648514	1297040	1297093	$2,42\ x10^{15}$
REGIONAL	0,3007	0,3007	226.859,5	2535,38	0	8	-648535	1297090	1297178	$2,43\ x10^{15}$
TIPO CONSTRUTIVO	1,52E-05	-6E-06	271.276,1	0,72	0,397411	1	-656972	1313949	1313976	$3,47\ x10^{15}$

A Área Construída Adquirida apresenta um R^2 de 0,604, indicando que em torno de 60% da variação na variável resposta pode ser explicada por esta variável, tornando-a relevante para a modelagem, e seu p-valor<0,05, (0), indica que a variável é altamente significativa, confirmando sua relevância na formação do preço dos imóveis.

O índice de Custo da Construção, ICC-BH, embora apresente um R^2 de 0,0079, indicando baixa representatividade na variável resposta, seu P-valor<0,05, (7,39E-84), é estatisticamente significante, e por ter uma correlação positiva com a variável resposta e ser um índice atrelado a construção de novos imóveis, este pode ser considerado um bom índice a ser considerado como componente de um modelo preditivo.

A Idade do Imóvel, embora apresente um R^2 de 0,0019, indicando baixa representatividade na variável resposta, seu P-valor<0, (8,64E-22), é estatisticamente significante. Essa significância pode contribuir no indicativo da depreciação ou obsolescência dos imóveis, podendo ter um leve impacto no preço

da variável resposta.

O Componente de Mão de Obra e o Componente de Material, que compõem o índice SINAPI, demonstram um R^2 de 0,0072 e 0,0067, respectivamente, indicando uma baixíssima representatividade na variável resposta. E apesar do seus P-valores serem menores que 0, (1,28E-76) e (6,09E-72), respectivamente, apresentando um certo nível de significância, ao levar em consideração a correlação destes índices com a variável resposta, apresentada na secção anterior, percebe-se que as correlações apresentaram valores baixos, tornando a contribuição destes índices em um modelo linear não relevante, podendo ser descartados em um modelo preditivo.

O índice de Inflação-BH demonstra um R^2 de 0,0005, e uma correlação com a variável resposta apresentada na secção anterior, muito fraca, e mesmo o P-valor sendo menor que 0, (5,47E-07), este índice tem um impacto sobre os preços dos imóveis quase irrelevante.

O índice de variação de aluguéis residenciais, IVAR-BH, apresenta um R^2 de 0,0065, que apesar de indicar uma baixa representatividade na variável resposta, ele possui um P-valor menor que 0, (9,43E-70), apresentando-se um certo nível de significância para um modelo linear. Analisando sua correlação com a variável resposta apresentada na secção anterior, e observando que é positiva, por estar ser um indicador que apesar de não estar diretamente relacionado com compra de imóveis, ele pode contribuir no preço dos imóveis, visto que pode ser observado como um indicador relacionado a habitação das famílias na cidade de belo Horizonte, podendo de forma indireta interferir na decisão de compra de um imóvel impactando no preço dos imóveis no mercado imobiliário de Belo Horizonte.

O Padrão de Acabamento demonstra um R^2 de 0,3013, apresentando cerca de 30% de representatividade na variável resposta, isto aliado ao P-valor = 0, indica relevância e significância desta variável com a variável resposta, sendo relevante em um modelo preditivo.

A variável Regional com o R^2 de 0,3007, representando 30% desta variável no preço do Valor de venda declarado do imóvel, aliado ao P-valor = 0, também sugere que a localização geográfica do imóvel é um fator relevante na formação do preço, e pode ser aplicado em um modelo preditivo.

Por se tratar de uma variável categórica que representa uma distinção estrutural

relevante entre os imóveis. Apesar do baixo R^2 de 1,52E-05 e do alto P-valor de 0,397411, maior que 0,05, indicarem estatisticamente que essa variável não explica significativamente a variação na variável resposta. A diferença na quantidade de dados de apartamentos e casas, 221.893 e 21.329, respectivamente, pode capturar efeitos que interagem com outras variáveis do modelo. Além disso, sua inclusão permite avaliar possíveis impactos indiretos quando combinada com fatores como Área Construída Adquirida, Idade do Imóvel e Padrão de Acabamento, contribuindo para uma modelagem mais representativa do mercado imobiliário.

Baseado no comportamento das variáveis quando aplicadas de forma individual, as variáveis Regional, Área Construída Adquirida, Tipo Construtivo, Padrão de Acabamento, Idade do Imóvel, e os índices ICC-BH e IVAR-BH foram escolhidas para comporem um modelo linear múltipla.

Tabela 10: Estatísticas do modelo linear aplicado as variáveis selecionadas.

Estatísticas do modelo linear aplicado as variáveis selecionadas				
Termos	**Coeficiente estimado**	**Erro Padrão da Estimativa**	**T-valor do modelo**	**P-valor**
(Intercept)	-208.817,93	6428,80	-32,48	6,5746E-229
REGIONALCENTRO-SUL	264.823,61	3308,94	80,03	0
REGIONALLESTE	80.537,27	3825,20	21,05	5,89997E-98
REGIONALNORDESTE	52.037,69	3259,57	15,96	3,18811E-57
REGIONALNOROESTE	45.505,91	3741,20	12,16	5,47267E-34
REGIONALNORTE	22.637,97	3482,07	6,50	8,04247E-11
REGIONALOESTE	73.236,20	2950,10	24,82	3,5718E-135
REGIONALPAMPULHA	68.675,91	2920,24	23,52	1,3671E-121
REGIONALVENDA NOVA	33.120,91	3322,32	9,97	2,19235E-23
Área.Construida.Adquirida	2.770,14	11,60	238,90	0
Tipo.Construtivocasa	-12.971,99	2656,82	-4,88	1,05075E-06
Padrao.AcabamentoPadrao Medio	-70.859,04	1630,81	-43,45	0
Padrao.AcabamentoPadrao Muito Bom	109.997,82	3814,80	28,83	2,9954E-181
Padrao.AcabamentoPadrao Muito Simples	-99.515,26	8760,57	-11,36	7,27889E-30
Padrao.AcabamentoPadrao Simples	-78.102,75	3294,73	-23,71	1,6812E-123
IdadeImóvel	-2.566,72	85,54	-30,01	5,668E-196
ICC_bh	218,89	10,27	21,31	2,8947E-100
IVAR_bh	828,59	95,51	8,68	4,23927E-18

No contexto da regressão linear, os coeficientes estimados para as variáveis independentes mostrados na Tabela 10 representam a variação média esperada na variável resposta (Valor Declarado) associada a uma unidade de alteração na

variável preditora, mantendo-se constantes todas as demais variáveis do modelo. Por exemplo, para a variável Área Construída Adquirida, o coeficiente de 2.770,14 indica que, para cada metro quadrado adicional de área construída adquirida, o valor do imóvel aumenta, em média, R$2.770,14. De forma análoga, a variável Idade Imóvel apresentou coeficiente negativo (-2.566,72), sugerindo que a cada ano adicional de idade do imóvel, ocorre uma depreciação média de R$2.566,72. No caso das variáveis categóricas como Regional, os coeficientes indicam um acréscimo médio no valor do imóvel em relação à região de referência, no caso do modelo a regional Barreiro, quando comparado a ela os imóveis situados na regional Centro-Sul apresentam valorização média de R$ 264.823,61, enquanto imóveis na regional Venda Nova apresentam uma valorização média de R$ 33.120,91. Para a variável Padrão de Acabamento, o coeficiente para imóveis com acabamento "Muito Bom" indica valorização média de R$ 109.997,82 em relação ao padrão de referência “Padrão Bom”. Por outro lado, padrões inferiores, como "Padrão Simples" e "Padrão Muito Simples", demonstraram depreciação média de R$ 78.102,75 e R$ 99.515,26, respectivamente.

Avaliando individualmente esses coeficientes percebe-se que a variável Regional apresentou coeficientes significativos para todas as categorias, com destaque para a regional Centro-Sul, que teve um coeficiente de R$264.823,61, indicando que imóveis localizados nessa região possuem, em média, um valor declarado significativamente superior ao das demais regionais. Em contraste, regiões como regional norte (R$22.637,97) e regional venda nova (R$33.120,91) possuem coeficientes positivos, mas de menor magnitude, sugerindo uma valorização menos expressiva.

A variável Área Construída Adquirida apresentou um coeficiente positivo de R$2.770,14, indicando que a cada metro quadrado adicional construído, o valor do imóvel tende a aumentar nesta proporcionalidade. Além disso, essa variável teve um alto nível de significância estatística (P-valor < 0,05), reforçando sua relevância no modelo.

Em relação a variável Tipo Construtivo, os resultados indicam que casas possuem, em média, um valor R$12.971,99 menor do que apartamento, mas por possuir um P-valor menor que 0,05, (1,05075E-06), observa-se sua significância no modelo linear.

A variável Padrão de Acabamento, apresentou impacto significativo sobre a variável

resposta. As classificações de Padrão Muito Bom possuem um acréscimo de R$109.997,82, indicando sua alta valorização, enquanto aqueles com classificados como Padrão Simples (R$-78.102,75) ou Padrão Muito Simples (R$-99.515,26) apresentam reduções expressivas no valor declarado, explicitando a importância da qualidade do acabamento na precificação dos imóveis.

A variável Idade do Imóvel teve um coeficiente negativo de R$-2.566,72, indicando que, a cada ano adicional de idade do imóvel, há uma redução média desse valor, refletindo a tendência de desvalorização ao longo do tempo, possivelmente estando relacionado a depreciação em virtude de problemas estruturais e necessidade de reformas com o passar do tempo.

As variáveis relacionadas a índices macroeconômicos também foram relevantes para o modelo. O ICC_bh apresentou um coeficiente positivo de R$218,89, enquanto o IVAR_bh teve um coeficiente de R$828,59, e ambos com P-valor menor que 0,05, 2,8947E-100 e 4,23927E-18, respectivamente, indicando uma significância estatística no modelo linear.

A análise de multicolinearidade do modelo linear, foi realizada pelos VIF valores (Tabela 11), que não indicaram problemas graves de correlação entre as variáveis, uma vez que todos os valores permaneceram abaixo de 5.

Tabela 11: Estatísticas do modelo linear aplicado as variáveis selecionadas.

VIF do modelo linear aplicado as variáveis selecionadas	
Variáveis	**GVIF**
REGIONAL	1,560681
Área.Construida.Adquirida	1,391671
Tipo.Construtivo	1,166174
Padrao.Acabamento	1,770327
IdadeImóvel	1,214167
ICC_bh	4,052696
IVAR_bh	4,031508

A qualidade do ajuste do modelo, pode ser melhor avaliada pelo R^2, e analisando o R^2 ajustado de 0,7485, nota-se que aproximadamente 74,85% da variabilidade do preço dos imóveis é explicada pelas variáveis incluídas na regressão.

Esses resultados demonstram que o modelo de regressão linear fornece uma explicação consistente para a precificação dos imóveis, destacando a importância da

localização, da área a ser adquirida do imóvel, do padrão de acabamento e a idade do imóvel como os principais fatores determinantes.

Mesmo com resultados consistentes, foi avaliada a possibilidade aplicar a técnica de Box-Cox à variável resposta com transformação logarítmica. Para isso, foi utilizada a função boxcox() do pacote MASS. A aplicação desta função à variável resposta permitiu encontrar um lambda de -0,22 que apesar de próximo de 0, não é suficientemente próximo para indicar que a transformação logarítmica traria ganhos significativos em um modelo linear com a variável resposta transformada, mantendo-se a escala original na regressão linear.

Apesar dos resultados já razoáveis obtidos através do modelo linear, para fins de comparação, as mesmas variáveis foram aplicadas em um modelo Random Forest, com os resultados apresentados na Tabela 12. O modelo Random Forest foi aplicado com um total de 100 árvores, selecionando duas variáveis aleatórias em cada divisão para determinar a melhor separação dos dados. Os resultados demonstraram um erro médio quadrático dos resíduos (MSR) de 13.536.444.883 e uma porcentagem da variância explicada de 81,61%, sendo este valor um pouco maior que o R^2 do modelo linear, mas como apresentado, o modelo Random Forest apresenta uma variabilidade maior quando comparada ao modelo linear.

Tabela 12: Estatísticas do modelo Random Forest aplicado as variáveis selecionadas.

Estatísticas do modelo Random Forest aplicado as variáveis selecionadas		
Variável	**%IncMSE**	**IncNodePurity**
REGIONAL	44,59	400024232155607
Área.Construida.Adquirida	82,52	1927246317498733
Tipo.Construtivo	24,33	37355563196198
Padrao.Acabamento	28,43	385657349742166
IdadeImóvel	47,82	138892274868994
ICC_bh	35,76	110408523577537
IVAR_bh	22,37	90247497058599

3.3 Avaliação das medidas de predição no modelo de regressão linear

A avaliação do modelo de regressão linear foi realizada comparando seu desempenho nas bases de treino e teste, utilizando métricas estatísticas como o

Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE), Coeficiente de Determinação (R^2), R^2 ajustado e o Erro Percentual Absoluto Médio (MAPE). Estas métricas aliadas a análise gráfica dos resultados Valores Preditos x Reais, revelam aspectos complementares quando avaliados conjuntamente.

Ao aplicar o modelo linear na base de testes e treino, os resultados das qualidades das predições são indicados na Tabela 13. Vê-se que o modelo gera predições com qualidade similar em ambos os conjuntos de dados. Os resultados obtidos demonstram que o modelo apresentou um R^2 de 0,75 na base de treino e 0,74 na base de teste, indicando que ele consegue explicar cerca de 74% da variação dos preços dos imóveis na base de teste. A proximidade desses valores sugere que o modelo generaliza bem e não apresenta sobre ajuste severo, já que sua capacidade explicativa se mantém estável em dados não utilizados no treinamento.

O MAE obtido foi R$91.260 na base de treino e R$92.902 na base de teste, indica em média, que a diferença absoluta entre os valores reais e os valores preditos pelo modelo está em trono de R$91.260. Já o MAPE de 23,04% na base de treino e 23,72% na base de teste, indicam em média, as previsões do modelo apresentam um erro de aproximadamente 23% em relação aos valores reais. A diferença relativamente pequena entre os valores desta métrica reforça a capacidade do modelo de prever novos dados com razoável precisão.

Entretanto, o Erro Quadrático Médio (MSE) apresentou um aumento significativo na base de teste (de R$18,5 bilhões para R$177,8 bilhões), revelando um impacto de outliers na performance do modelo, uma vez que essa métrica penaliza erros maiores de forma quadrática. Isso também se reflete no RMSE, que passou de R$136.022 na base de treino para R$139.259 na base de teste, evidenciando um leve aumento na dispersão dos erros.

Tabela 13: Qualidade das predições do modelo de regressão linear aplicado nas bases de teste e treino.

Qualidade das predições do modelo linear aplicado a base de treino e teste						
TIPO_Reg	**MAE**	**MSE**	**RMSE**	**R_Quad**	**R_Quad_Ajus**	**MAPE**
Reg Linear base treino	91260,29	18502035837	136022,19	0,75	0,75	23,04
Reg Linear base teste	92902,26	1,77815E+11	139259,66	0,74	0,74	23,72

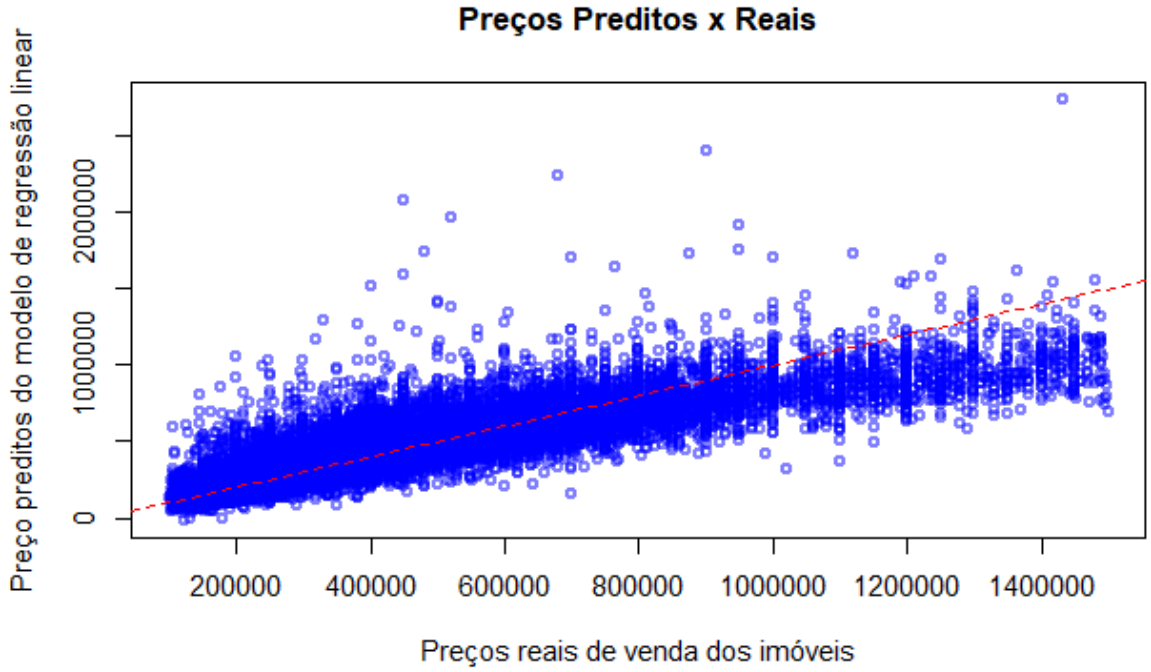
A relação entre os valores preditos pelo modelo na base de teste (eixo horizontal) e

os valores reais dos preços de venda dos imóveis (eixo vertical) são mostrados Figura 10. A reta pontilhada vermelha representa a linha ideal onde os valores preditos seriam exatamente iguais aos valores reais (y=x). A distribuição dos pontos indica que o modelo consegue captar a tendência geral do mercado imobiliário, uma vez que os valores preditos seguem uma tendência ascendente à medida que os preços reais aumentam. No entanto, observa-se uma dispersão significativa, principalmente para imóveis de maior valor, o que sugere dificuldades do modelo em prever corretamente preços elevados.

Para imóveis com valores inferiores a aproximadamente R$600.000, o modelo apresenta um ajuste mais preciso, com menor dispersão em torno da linha de referência. No entanto, à medida que os preços dos imóveis aumentam, a dispersão dos pontos cresce consideravelmente, indicando maiores erros na previsão. Especificamente, a partir de valores superiores a R$1.200.000, o modelo frequentemente subestima os preços, pois muitos pontos estão posicionados abaixo da reta ideal. Essa subestimação pode estar associada a limitações do modelo linear em capturar variações estruturais mais complexas como o padrão de acabamento desses apartamentos e a área construída adquirida na compra deles, que podem apresentar uma variabilidade significante na predição do modelo, interferindo nos resultados.

Esses resultados corroboram com o percentual de 75% do R^2, que apesar de ser um resultado bom para compreendera variabilidade do modelo, outros 25% ainda são comprometidos, podendo obter resultados melhores na aplicação dos dados em outro modelo de predição.

Figura 10: Valor Predito x Valor Real para modelo de regressão linear quando aplicado no conjunto de testes.



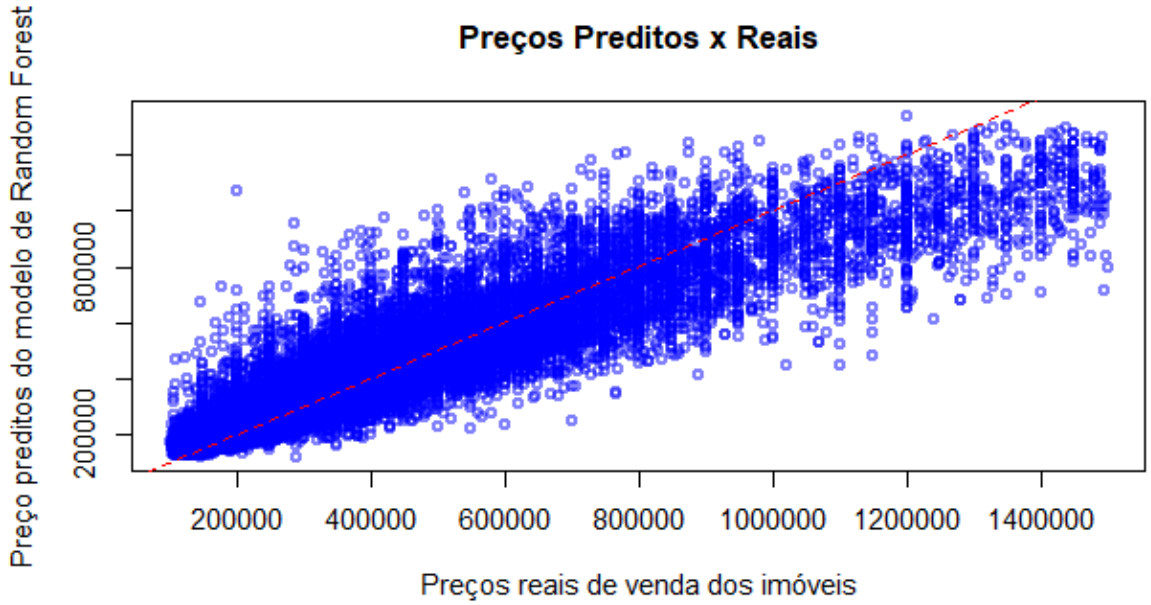
3.4 Avaliação das medidas de predição no modelo Random Forest

Uma análise similar sobre a qualidade preditiva foi realizada para o modelo Random Forest e os resultados são exibidos na Tabela 14 e Figura 11.

Tabela 14: Qualidade das predições do modelo Random Forest aplicado a base de treino e teste.

Qualidade das predições do modelo Random Forest aplicado a base de treino e teste						
TIPO_Reg	**MAE**	**MSE**	**RMSE**	**R_Quad**	**R_Quad_Ajus**	**MAPE**
Random Forest base treino	53383,76	6,53E+09	80794,17	0,82	0,82	12,88
Random Forest base teste	76336,52	1,34E+10	115937,1	0,82	0,82	18,78

Figura 11: Valor Predito x Valor Real para modelo Random Forest quando aplicado no conjunto de testes.



O MAE, que representa a média dos erros absolutos entre os valores previstos e os valores reais, fornece uma medida da magnitude dos erros do modelo, sem considerar a direção do erro. Os valores obtidos para o MAE na base de treino e teste, R$53.383,76 e R$76.336,52, respectivamente, indicam uma discrepância de aproximadamente R$23.000. Esta variação no erro entre treino e teste pode ser um indicativo de sobre ajuste, visto que, o modelo aprendeu muito bem os padrões da base de treino, mas tem dificuldades em generalizar para dados da base de testes.

Avaliando o MSE, e o RMSE, que são indicadores que medem a magnitude dos erros de previsão, ou seja, a discrepância entre os valores reais e os valores preditos pelo modelo das bases de treino e teste. Percebe-se diferenças destes valores entre esses parâmetros nas bases de treino e teste. Sendo R$6.530.000.000 para o MSE da base de treino e R$13.400.000.000 para o MSE da base de testes, assim como R$80.794,17 para o RMSE da base de treino e R$115.937,1 para o RMSE da base de teste. Estes resultados apresentam uma diferença de R$6.870.000.000 para o MSE e aproximadamente R$35.000 para o RMSE entre as bases respectivamente. Evidenciando que este aumento substancial destes índices entre as bases sugere que o modelo tem dificuldades em lidar com erros grandes em dados na base de teste, reforçando a ideia de sobre ajuste, onde o modelo se ajusta bem aos dados de treino, mas tem a performance degrada quando confrontado com a base de dados de testes.

O MAPE, que quantifica a precisão do modelo em termos percentuais, indicando o

quanto as predições do modelo estão distantes dos valores reais, apresentaram para a base de treino um valor de 12,88% e para a base de teste um valor de 18,78%, indicando uma precisão em torno de apenas 6% menor na base de teste.

O R^2, sendo o indicador que mede a proporção da variabilidade total dos dados que é explicada pelo modelo, assim como o R^2 ajustado que leva em consideração o número de variáveis preditoras e ajusta a medida para evitar a superestimação do desempenho do modelo, apresentaram o mesmo valor de 0,82 em ambas as bases, sugerindo que o modelo consegue explicar uma boa parte da variabilidade dos dados, tanto no conjunto de treino quanto no de teste. E mesmo com outras métricas como MAE, MSE, RMSE e o MAPE indicando que o modelo pode capturar bem a estrutura geral dos dados, na base de teste, mas não necessariamente na base de treino, o R^2 representando cerca de 82% da variabilidade dos dados da variável resposta sobre a variável dependente, apresenta um bom indicativo da aplicação do modelo Random Forest tanto na base de treino quanto de testes.

Avaliando Figura 5, que mostra relação entre os preços reais de venda dos imóveis e os preços estimados pelo modelo Random Forest, através de um gráfico de dispersão, em que, no eixo X, encontram-se os valores reais dos imóveis, e no eixo Y estão os valores preditos pelo modelo Random Forest com a reta pontilhada vermelha representando a linha ideal onde os valores preditos seriam exatamente iguais aos valores reais (y=x), percebe-se que a distribuição dos pontos do modelo Random Forest consegue captar a tendência geral do mercado imobiliário, uma vez que os valores preditos seguem uma tendência ascendente à medida que os preços reais aumentam. No entanto, observa-se uma dispersão significativa, principalmente para imóveis de maior valor, o que sugere dificuldades do modelo em prever corretamente preços elevados.

Para imóveis com valores entre R$200.000 e R$850.000, o modelo apresenta um ajuste mais preciso, com menor dispersão em torno da linha de referência. No entanto, à medida que os preços dos imóveis aumentam, a dispersão dos pontos cresce consideravelmente, indicando maiores erros na previsão. Especificamente, a partir de valores superiores a R$900.000, o modelo frequentemente subestima os preços, pois muitos pontos estão posicionados abaixo da reta ideal. Essa subestimação pode estar associada a limitações do modelo linear em capturar variações estruturais mais complexas como o padrão de acabamento de

apartamentos a partir desta faixa de valor, assim como a área construída adquirida na compra deles. Podendo estes fatores apresentarem uma variabilidade significativa na predição do modelo Random Forest, interferindo nos resultados deste modelo.

Apesar disso, através da análise gráfica aliada com a análise dos dados sobre a qualidade preditiva do modelo Random Forest, verifica-se que para pelo menos 82% da variabilidade dos dados da variável resposta, este modelo apresenta um bom resultado para análise preditiva da precificação dos imóveis na cidade de Belo Horizonte.

3.5 Comparação entre os modelos ajustados

Para facilitar a comparação entre os dois modelos aplicados, regressão linear e Random Forest, os resultados das medidas de qualidade preditiva para ambos são apresentadas na Tabela 15.

Tabela 15: Tabela comparativa dos resultados preditivos dos dois modelos para o conjunto de treino.

Tabela comparativa dos resultados preditivos dos dois modelos para o conjunto de treino						
Tipo de regressão	**MAE**	**MSE**	**RMSE**	**R_Quad**	**R_Quad_Ajus**	**MAPE**
Reg Linear base treino	91.260,29	1,85E+10	136.022,2	0,75	0,75	23,04
Random Forest base treino	53.383,76	6,53E+09	80.794,17	0,82	0,82	12,88
Reg Linear base teste	92.902,26	1,77815E+11	139.259,66	0,74	0,74	23,72
Random Forest base teste	76.336,52	1,34E+10	115.937,1	0,82	0,82	18,78

O MAE, que mede o erro médio absoluto entre as previsões e os valores reais, apresentou uma diferença significativa entre os dois modelos, tanto na base de treino quanto na base de teste. O MAE na base de treino, do Random Forest foi de R$53.383,76, enquanto o da Regressão Linear foi de R$91.260,29, já na base de teste foi de R$92.92,26 da Regressão Linear e R$76.336,52 do Random Forest. Em ambas as bases, esse resultado sugere que o modelo Random Forest é capaz de prever os preços dos imóveis com um erro médio inferior, indicando uma maior precisão nas suas previsões em comparação com a Regressão Linear.

O MSE e o RMSE fornecem uma medida de erro mais sensível a grandes discrepâncias entre as previsões e os valores reais. Na base de treino o MSE do Random Forest de R$6.527.697.962, é significativamente menor do que o MSE da

Regressão Linear de R$18.502.035.837, enquanto na base de teste o MSE do Random Forest é R$13.441.402.814, também sendo significativamente menor do que o da Regressão Linear de R$177.814.930.211. Já o RMSE do Random Forest na base de treino é de R$80.794,17, contra R$136.022,19 da Regressão Linear, e na base de teste, este indicador no modelo Random Forest apresentou um valor R$115.937,1, e da Regressão linear R$139.259,66, evidenciando uma magnitude de erro menor no Random Forest em ambas as bases, que pode ser interpretado como um modelo mais preciso e eficiente na previsão dos preços dos imóveis.

O Coeficiente de Determinação R^2 e R^2 ajustado indicam a proporção da variação dos dados explicada pelo modelo, tanto na base de treino como na de testes, ambos os índices apresentaram valores mais elevados para o Random Forest (0,82), quando comparado com os valores destes índices para a Regressão Linear (0,75 na base de treino e 0,74 na base de teste). Esse aumento no R^2 sugere que o modelo de Random Forest tem uma capacidade explicativa superior, sendo mais eficaz na explicação da variabilidade dos preços dos imóveis indicando que o modelo é mais robusto e tem um melhor desempenho ajustado para o número de variáveis preditoras, quando comparado com o modelo linear.

O MAPE, que indica o erro médio percentual das previsões, também foi mais baixo em ambas as bases. Na base de treino o Random Forest, apresentou um valor de 12,88% vide os 23,04% da Regressão Linear, já na base de testes, o Random Forest apresentou um valor de 18,78% contra 23,72% da Regressão Linear, estes valores mais baixos do MAPE sugerem que o modelo Random Forest é mais preciso em termos relativos, cometendo erros menores na precificação dos imóveis, se comparado ao modelo linear.

Observando os resultados entre os modelos, o modelo de predição aplicado com Random Forest supera a Regressão Linear em todas as métricas de avaliação, apresentando um erro médio menor, um MSE e RMSE menor, uma maior capacidade explicativa, com um R^2 melhor e maior precisão relativa (MAPE) em comparação com a Regressão Linear. Esses resultados indicam que, tanto para a base de treino quanto a de teste, o modelo Random Forest é mais eficiente e preciso na previsão de preços dos imóveis, tornando-se uma escolha mais robusta para uma tarefa preditiva.

4. CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma precificação de imóveis na cidade de Belo Horizonte, utilizando dados compreendidos entre 2019 e 2024. Essa precificação é significativa tanto para corretores quanto para imobiliárias atuantes no mercado local. Com a aplicação dos modelos de regressão linear e Random Forest na base de dados, os resultados indicaram boa qualidade preditiva segundo o coeficiente de determinação ajustado sendo igual a 82% na explicação da variabilidade dos preços no modelo Random Forest. Com esse resultado, a modelagem proporciona uma estimativa realista e alinhada com os preços dos imóveis vendidos em Belo Horizonte.

Apesar do bom desempenho obtido, ainda há espaço para aprimoramentos em pesquisas futuras. Ao alinhar a base de dados pública dos registros de valores de ITBI pagos com outras bases que contenham informações mais detalhadas sobre os imóveis, como quantidade de quartos, banheiros e vagas de garagem, há a possibilidade de aumentar a precisão na precificação dos imóveis. Além disso, a incorporação de técnicas de predição mais avançadas pode contribuir para um ajuste mais adequado dos preços, especialmente em imóveis de alto valor agregado.

A implementação dessas melhorias metodológicas em estudos futuros pode resultar em modelos com maior poder explicativo, ampliando suas aplicações práticas e contribuindo para uma precificação mais precisa no mercado imobiliário da capital mineira.

5. REFERÊNCIAS

BREIMAN, L. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001

Diário do Comércio. Economia. **Volume de imóveis vendidos em Belo Horizonte cresce 40% e fortalece setor.** Disponível em: **https://diariodocomercio.com.br/economia/volume-imoveis-vendidos-belo-horizonte-cresce/**. Acesso em: 20 nov. 2024.

FGV-IBRE. **ICC-BH, IGP-M, IVAR-BH**. Disponível em: https://extra-ibre.fgv.br/IBRE/sitefgvdados/consulta.aspx. Acesso em 13 jan. 2025.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. Multivariate Data Analysis. Cengage Learning, 2019.

IBGE. **INFLAÇÃO**. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?t=series-historicas&utm_source=landing&utm_medium=explica&utm_campaign=inflacao#plano-real-mes. Acesso em 13 jan. 2025.

IPEAD. **IPCA-BH**. Disponível em: https://www.ipead.face.ufmg.br/site/publicacoes/indicesPrecos. Acesso em 13 jan. 2025.

LIAW, Andy; WIENER, Matthew. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.7-1, 2023. Disponível em: https://cran.r-project.org/web/packages/randomForest/. Acesso em: 28 mar. 2025.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. Applied Linear Statistical Models. McGraw-Hill/Irwin, 2013.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to Linear Regression Analysis. John Wiley & Sons, 2021.

Neri, Evandro. 1843. **MODELO PREDITIVO DO PREÇO DE VENDA DE APARTAMENTOS EM BELO HORIZONTE UTILIZANDO RANDOM FOREST.** Disponível em: **https://repositorio.ufmg.br/community-list**. Acesso em 02 nov. 2024

Paixão, Luiz Andrés Ribeiro. v. 19 n. 1 (2015). **ÍNDICE DE PREÇOS HEDÔNICOS PARA IMÓVEIS: UMA ANÁLISE PARA O MUNICÍPIO DE BELO HORIZONTE**. Disponível em: **https://www.revistas.usp.br/ecoa/**. Acesso em: 10 nov. 2024.

Portal de dados abertos da prefeitura de Belo Horizonte. **01/2008 a 05/2024 - ITBI Relatórios**, https://ckan.pbh.gov.br/dataset/itbi-relatorios. Acesso 01 ago. 2024

Prefeitura de Belo Horizonte, 2019.**RELAÇÃO DE BAIRROS, REGIONAL E TERRITÓRIOS DE GESTÃO COMPARTILHADA**. Disponível em: https://prefeitura.pbh.gov.br/sites/default/files/estrutura-de-governo/cultura/2019/COMUC/Rela%C3%A7%C3%A3o%20de%20bairro%2C%20regional%20e%20territ%C3%B3rios.pdf. Acesso 15 jan. 2025.

SIDRA. **INPC, SINAPI**. Disponível em: https://sidra.ibge.gov.br/home/inpc/brasil. Acesso em 13 jan. 2025.

TUKEY, J. W. Exploratory Data Analysis. Addison-Wesley, 1977.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, v. 30, n. 1, p. 79-82, 2005.