RK&MS Data Science

Consultoria de dados

Marcos Vinicius Sokabe Ribeiro[1]

Rafael Kobayashi[2]

Estudo de Caso: Apartamento em Perdizes, 56m²

Pesquisa de mercado

São Paulo

2025

[1]Marcos Vinicius Sokabe Ribeiro é Jornalista pela Pontifícia Universidade Católica de São Paulo e Cientista de Dados pela Faculdade de Tecnologia do Estado de São Paulo, campus Cotia.
[2]Rafael Kobayashi é Cientista de Dados pela Faculdade de Tecnologia do Estado de São Paulo, campus Cotia.

Estudo de Caso: Apartamento em Perdizes, 56m²

Estudo de caso de imóvel com fins corporativos apresentado à EMPRESA X, como documento de pesquisa e análise de dados seguindo as normas ABNT e diretrizes corporativas do NBS e Serviços.

São Paulo

2025

VISTO DE AUTORIA

_____

_____

> quanto mais um estudioso estiver aparentemente dispersando sua investigação entre os fatos [...] melhor poderá observar o que está acontecendo na sociedade que estuda. Poderá formular suas hipóteses e obter seus dados em qualquer setor da vida com vantagem.
>
> (BENEDICT, Ruth, 1976, pag.18-19)

RESUMO

RK&MS DATA SCIENCE. **Estudo de Caso: Apartamento em Perdizes, 56m²**

Este estudo apresenta uma análise quantitativa do mercado imobiliário de apartamentos de 56m² no bairro de Perdizes, São Paulo, utilizando técnicas de ciência de dados e web scraping para coleta e processamento de dados. A pesquisa teve como objetivo determinar o valor de mercado e identificar padrões de precificação através de métodos estatísticos aplicados a dados coletados automaticamente de plataformas imobiliárias.

Palavras-chave: perdizes; são paulo; estatística; ciência de dados; estimativa.

ABSTRACT

RK&MS DATA SCIENCE. **Case Study: Apartment in Perdizes, 56m²**

This study presents a quantitative analysis of the real estate market for 56m² apartments in the Perdizes neighborhood, São Paulo, using data science techniques and web scraping for data collection and processing. The research aimed to determine the market value and identify pricing patterns through statistical methods applied to data automatically collected from real estate platforms.

Key-words: perdizes; são paulo; estatística; ciência de dados; estimativa.

LISTA DE FIGURAS

Figura 1 - Exibição parcial de código de mineração de dados 15

Figura 2 - Exibição parcial de código de saneamento de dados 17

Figura 3 - Histograma mostra a distribuição de dados de imóveis da população crua, sem tratamento 19

Figura 4 - Gráfico de densidade probabilística revela alta concentração de valores em intervalos iniciais da população crua 20

Figura 5 - Boxplot mostra os valores considerados outliers da população crua .. 20

Figura 6 - QQ-Plot atesta as observações de gráficos anteriores sobre a população crua não seguir uma distribuição normal 21

Figura 7 - Gráfico de Histograma e Densidade mostra a distribuição de dados após o processamento e normalização 22

Figura 8 - Gráfico de Histograma e Densidade exibe a distribuição de M2 na amostra; valores maiores que 60M2 servem de prova à inferência negativa de metragem como determinador de preço chamada de economia de escala 23

Figura 9 - Gráfico de Dispersão exibe o espaço amostral dentro de uma distribuição próxima à normal 24

Figura 10 - Matriz de Correlação entre as variáveis quantitativas do estudo ... 25

SUMÁRIO

TEXTO PRELIMINAR 10
1 INTRODUÇÃO 13
2 OBJETIVO 14
3 MATERIAIS 15

3.1 ALGORITMO DE WEBSCRAPING 15
3.2 ALGORITMO DE PRÉ-PROCESSAMENTO 16

4 METODOLOGIA 19

4.1 AMOSTRAGEM 19
4.2 ANÁLISE EXPLORATÓRIA 23

5 RESULTADOS 28

5.1 Análise Descritiva da Amostra Final 28
5.2 Estimativa de Preço e Intervalo de Confiança 28
5.3 Análise de Fatores Determinantes 29
5.4 Validação Estatística 29

6 CONCLUSÃO 30
REFERÊNCIAS 32
ANEXO A -- Dados Brutos da Amostra 33

TEXTO PRELIMINAR

Há diferenças notáveis entre avaliações de peritos técnicos (corretores de imóveis) e estimativas de cientistas de dados. Cada qual, consagrada em seu próprio saber, apresenta limitações, potencialidades e intersecções. Suas diferenças, entretanto, não podem ser reduzidas a métodos superiores ou inferiores, assim como suas aproximações não podem apagar suas singularidades teóricas. Quando compreendidas como abordagens complementares que refletem diferentes epistemologias, o conhecimento do setor imobiliário se desenvolve.

Os corretores, através de seu conhecimento especializado, oferecem uma visão local de conceitos ligados à avaliação de imóveis. Suas avaliações são potencializadas por anos de experiência prática no setor e uma espécie de know-how que, por sua natureza, não exige rigor científico. Nesse trabalho, há o pressuposto de que a experiência do corretor é um elemento que enriquece o processo de avaliação. Mas se os corretores pensam e operam a partir de visões localizadas em que a própria subjetividade do autor é bem-vinda, cientistas de dados têm, em seu tratado teórico, a busca pela objetividade e eliminação de vieses, culminando em protocolos rigorosos e uma tendência à visão de grande contingência.

Esta distinção epistemológica fundamental revela tensões produtivas que merecem aprofundamento. A valorização da subjetividade pelos corretores representa o argumento de que o mercado imobiliário é constituído por elementos irredutíveis à quantificação pura. Os cientistas de dados, no entanto, podem extrair padrões e insights dos dados que não são facilmente perceptíveis por uma avaliação estritamente humana. Suas estimativas são objetivas e baseadas em evidências, oferecendo uma abordagem mais sistemática para a tomada de decisões.

Aqui, vale apontar uma diferença semântica que esclarece a distinção entre o fazer específico de cada área de conhecimento em relação à precificação de ativos imobiliários: enquanto os corretores utilizam o termo ‘avaliação’ que, em sua etimologia, denota significados mais compostos e subjetivos, os cientistas de dados operam a partir do termo ‘estimativa’, com o sentido de resultados balizados por probabilidade e teste de hipóteses.

A aplicação do método científico na ciência de dados para precificação imobiliária manifesta-se através de protocolos estruturados que buscam transformar

a estimativa de imóveis em um processo empiricamente verificável e metodologicamente replicável. A distinção entre os métodos amostrais empregados por corretores e cientistas de dados revela diferenças fundamentais na construção do conhecimento sobre precificação imobiliária, com implicações diretas para a validade e generalização dos resultados obtidos.

A diferença mais evidente entre as abordagens reside no volume de dados analisados. Corretores baseiam suas avaliações em amostras reduzidas, tipicamente algumas dezenas de imóveis similares que conseguem recordar ou acessar em suas bases pessoais. Esta limitação amostral resulta em menor variabilidade nas estimativas e menor confiabilidade estatística das conclusões. Cientistas de dados trabalham com amostragens grandes obtidas por mineração de dados parametrizada. Este volume permite análises estatisticamente robustas, variabilidade real de mercado e identificação de padrões que seriam imperceptíveis em amostras menores. A lei dos grandes números garante que as médias amostrais convergem para os valores populacionais verdadeiros.

O processo de seleção de imóveis comparáveis pelos corretores introduz subjetividade significativa. A escolha de quais propriedades consideram "similares" depende de julgamentos pessoais sobre quais características são relevantes e como ponderá-las. Esta seleção subjetiva pode gerar inconsistências: diferentes corretores avaliando o mesmo imóvel podem escolher conjuntos distintos de comparáveis, resultando em estimativas divergentes. A ausência de critérios padronizados torna o processo pouco replicável. A ciência de dados emprega critérios objetivos e padronizados para definir similaridade entre imóveis. A composição da amostra considera um conjunto de variáveis preliminarmente definido conforme à necessidade da pesquisa. Há uma relação fina entre tamanho da amostra e quantidade de variáveis em jogo.

Amostras pequenas e selecionadas subjetivamente podem produzir estimativas precisas quando o conhecimento local do corretor é acurado, mas também amplificam erros quando os comparáveis escolhidos não são genuinamente representativos. A falta de padronização torna difícil avaliar sistematicamente a qualidade das estimativas. A abordagem de ciência de dados, com grandes amostras e critérios objetivos, tende a produzir estimativas mais consistentes e estatisticamente confiáveis, reduzindo a influência de casos atípicos e vieses de

seleção individual. A representatividade sistemática permite generalização mais segura dos resultados para o mercado como um todo.

1. INTRODUÇÃO

Este relatório é um estudo de precificação de um apartamento em Perdizes, São Paulo/SP, e tem como base o Método Comparativo de Mercado (DANTAS, 2012). Ele se baseia em conhecimentos multidisciplinares da ciência de dados, como tecnologia, pesquisa e estatística, e segue o formato recomendado pela ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2011).

Seu propósito é estimar o valor de um ativo imobiliário através dos protocolos consagrados na pesquisa por amostragem e inferência estatística (BUSSAB; MORETTIN, 2017). Por sua natureza, esse documento segue os princípios fundamentais do método científico: reprodutibilidade, falseabilidade e ciclismo (POPPER, 2013), extrapolando seu caráter de uso localizado no meio corporativo, a fim de contribuir com a produção de conhecimento do mercado imobiliário brasileiro.

A amostragem é produto de um algoritmo de webscraping construído na linguagem Python (MITCHELL, 2018). Ele parametriza a coleta de dados conforme as características do objeto de estudo. A fonte de pesquisa é a ZAP Imóveis, o maior portal imobiliário online do Brasil, garantindo tamanho, representatividade e reprodutibilidade (HAIR et al., 2019). As variáveis selecionadas neste estudo foram pensadas para equilibrar a noção de tamanho amostral e generalização de dados.

O tratamento de dados é feito via ETL em Python, com pré-processamento, identificação de outliers e análise de metadados (KIMBALL; ROSS, 2013). Medidas de tendência central compõem o arsenal estatístico descritivo, enquanto técnicas de regressão linear e análise multivariada fundamentam a modelagem preditiva (FIELD, 2018). A metodologia adotada permite não apenas a avaliação do imóvel específico, mas também a identificação de padrões de mercado, tendências de valorização e fatores determinantes de preço na região de Perdizes (GONZALEZ; FORMOSO, 2006). Os resultados obtidos são apresentados com intervalos de confiança estatística, proporcionando transparência e robustez técnica às conclusões (TRIOLA, 2017).

Este documento representa, portanto, uma aplicação prática da ciência de dados no setor imobiliário, contribuindo para a democratização do conhecimento técnico e para o aprimoramento das práticas de avaliação no mercado brasileiro (PROVOST; FAWCETT, 2013).

2. OBJETIVO

Este estudo tem como objetivo principal desenvolver um modelo de precificação imobiliária científico e reprodutível para estimar o valor de um apartamento no bairro de Perdizes, São Paulo/SP, utilizando o Método Comparativo de Mercado apoiado em técnicas avançadas de ciência de dados. A pesquisa busca implementar um sistema automatizado de coleta de dados através de webscraping em Python, parametrizado conforme as características específicas do imóvel em análise, utilizando a plataforma ZAP Imóveis como fonte representativa do mercado brasileiro.

O trabalho pretende estabelecer um pipeline completo de tratamento de dados via processo ETL, contemplando desde o pré-processamento até a identificação de outliers e análise de metadados, garantindo assim a qualidade e confiabilidade da base informacional. Através da aplicação de técnicas estatísticas descritivas e inferenciais, incluindo medidas de tendência central, regressão linear e análise multivariada, objetiva-se fundamentar solidamente a modelagem preditiva de preços imobiliários.

A validação rigorosa dos modelos desenvolvidos constitui aspecto central da pesquisa, utilizando métricas de performance estatística e análise de resíduos para assegurar a precisão das estimativas. Simultaneamente, busca-se identificar padrões de mercado e fatores determinantes na precificação da região de Perdizes, analisando tendências de valorização e variáveis que exercem influência significativa sobre os valores imobiliários.

Os resultados serão apresentados com intervalos de confiança estatística, proporcionando transparência metodológica e robustez técnica às conclusões, em consonância com os princípios científicos de reprodutibilidade e falseabilidade. Desta forma, o estudo pretende contribuir para a democratização do conhecimento técnico no setor imobiliário brasileiro, oferecendo uma metodologia científica replicável que pode ser aplicada em diferentes contextos geográficos e temporais, ampliando assim o conhecimento sobre práticas de avaliação no mercado nacional.

3. MATERIAIS

3.1 ALGORITMO DE WEBSCRAPING

O material utilizado para a obtenção da base de dados consistiu em um webscraper, uma ferramenta de software projetada para a coleta automatizada de informações disponíveis em páginas da internet. Esse sistema foi desenvolvido na linguagem de programação Python e opera em duas etapas principais: a coleta e o tratamento dos dados.

A fase de coleta emprega uma abordagem de navegação e extração de dados inspirada no comportamento humano. Em vez de simplesmente baixar o conteúdo da página, o webscraper utiliza um navegador controlado de forma automatizada. Essa abordagem simula ações de um usuário real, como a navegação entre as páginas, o movimento do cursor, e a rolagem da página. Isso garante que a ferramenta possa acessar conteúdos dinâmicos, que são carregados à medida que o usuário interage com a página.

```
import pandas as pd
import re
import numpy as np
import warnings
from scipy.stats import zscore
import time
import random
import undetected_chromedriver as uc
from fake_useragent import UserAgent
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import os
from IPython.display import clear_output, display
from bs4 import BeautifulSoup
try:
    import matplotlib.pyplot as plt
    import seaborn as sns
    PLOTTING_AVAILABLE = True
except:
```

Figura 1. Exibição parcial de código de mineração de dados

Após a coleta, os dados são processados. Cada anúncio extraído é analisado e as informações de interesse, como preço, área em metros quadrados, localização,

número de quartos, banheiros e vagas de garagem são identificadas e extraídas. O sistema conta com mecanismos de busca flexíveis para garantir que a maior quantidade de dados seja capturada. Além disso, a etapa de tratamento de dados inclui a remoção de duplicatas e a identificação de discrepâncias, ou outliers, que podem distorcer a análise estatística. Para lidar com esses outliers, o sistema aplica uma abordagem iterativa que combina os métodos do Intervalo Interquartil (IQR) e do Z-Score. A escolha entre substituir os valores discrepantes pela média ou pela mediana é feita de forma adaptativa, com base no coeficiente de variação dos dados, assegurando uma limpeza de dados mais robusta.

Finalmente, a base de dados resultante, limpa e estruturada, é analisada estatisticamente. Medidas de tendência central, como média e mediana, e de dispersão, como o coeficiente de variação, são calculadas e visualizadas em gráficos. A média ponderada por intervalos de preços também é calculada para fornecer uma estimativa mais precisa do valor do metro quadrado na região estudada. Todo o processo é automatizado e resulta em um arquivo de dados pronto para ser utilizado em modelos preditivos e análises estatísticas mais aprofundadas.

3.2 ALGORITMO DE PRÉ-PROCESSAMENTO

A análise dos dados do mercado imobiliário foi conduzida por meio de um sistema automatizado, desenvolvido em Python, que integra diversas etapas do processo de Análise Exploratória de Dados (AED) e limpeza de dados. A principal ferramenta utilizada para o processamento foi a biblioteca Pandas, conhecida por sua eficiência no manuseio e manipulação de conjuntos de dados. A visualização dos resultados foi realizada com o auxílio das bibliotecas Matplotlib e Seaborn, que permitiram a geração de gráficos para uma melhor compreensão da distribuição e das relações entre as variáveis.

```python
def remover_duplicadas(df):
    """Remove linhas duplicadas do DataFrame"""
    if df is None or df.empty:
        return df

    print(f"\nTotal de linhas antes da remoção de duplicadas: {len(df)}")

    # Remove duplicatas baseado em todas as colunas
    df_sem_duplicatas = df.drop_duplicates()

    # Se você quiser remover duplicatas apenas baseado em colunas específicas:
    # df_sem_duplicatas = df.drop_duplicates(subset=['Endereco', 'M2', 'Preco'])

    duplicatas_removidas = len(df) - len(df_sem_duplicatas)
    print(f"Duplicatas removidas: {duplicatas_removidas}")
    print(f"Total de linhas após remoção de duplicadas: {len(df_sem_duplicatas)}")

    return df_sem_duplicatas
```

Figura 2. Exibição parcial de código de saneamento de dados

O fluxo de processamento de dados pode ser detalhado nas seguintes etapas:

A) Carregamento e Inspeção Inicial

O processo inicia com o carregamento do arquivo de dados no formato .csv. Após o carregamento, é realizada uma inspeção inicial para verificar a integridade e a estrutura do conjunto de dados. Esta etapa envolve a exibição de informações fundamentais, como o número de linhas e colunas, os tipos de dados de cada variável e a presença de valores ausentes (missing values). A visualização das primeiras linhas e das estatísticas descritivas básicas, como média, desvio padrão e quartis, fornece um panorama inicial das características dos dados, permitindo a identificação de possíveis problemas antes da limpeza.

B) Limpeza de Dados

O passo seguinte concentra-se na limpeza e preparação dos dados para análise. Primeiramente, é executada a remoção de duplicatas para assegurar que cada registro na base de dados represente um imóvel único, eliminando redundâncias que poderiam distorcer a análise estatística.

Em seguida, o sistema realiza uma robusta substituição de outliers, que são valores extremos que se desviam significativamente do restante dos dados. Para isso, é empregado um método iterativo que combina duas técnicas estatísticas: o Método do Intervalo Interquartil (IQR) e o Método do Z-Score. A escolha entre substituir um outlier pela média ou pela mediana do conjunto de dados é feita de

forma adaptativa, baseando-se no Coeficiente de Variação. Se a variabilidade dos dados for baixa (ou seja, o coeficiente for menor que 40%), a média é utilizada por ser um estimador mais eficiente; caso contrário, a mediana é preferida por ser mais resistente à influência de valores extremos. Esse processo iterativo é repetido até que a distribuição dos dados se estabilize, garantindo uma base limpa e representativa do mercado.

C) Análise Exploratória e Estatísticas

Com os dados limpos, o sistema prossegue para a análise. As estatísticas descritivas da variável de interesse (e.g., preço por metro quadrado) são calculadas novamente, exibindo a média, mediana, moda, desvio padrão e quartis. Essas informações fornecem uma visão precisa das características da amostra após a remoção de valores atípicos.

D) Dataviz

Para complementar a análise, são gerados múltiplos gráficos:
Histograma e Boxplot: permitem a visualização da distribuição de frequência dos dados e a identificação visual de dispersões e concentrações. Gráfico de Densidade (KDE): apresenta uma estimativa da distribuição de probabilidade da variável. Q-Q Plot: avalia a normalidade da distribuição dos dados, comparando-a com uma distribuição normal teórica.

Além disso, a análise de correlação é executada para quantificar a relação linear entre as variáveis numéricas do dataset, sendo visualizada através de uma Matriz de Correlação. Essa etapa é crucial para entender como variáveis como área, número de quartos e vagas de garagem se relacionam com o preço do imóvel, oferecendo insights valiosos para a modelagem preditiva.

4. METODOLOGIA

4.1 AMOSTRAGEM

A coleta de dados foi realizada através de um webscraper parametrizado para buscar anúncios de apartamentos no portal Zap Imóveis. O universo da pesquisa foi definido para a cidade de São Paulo, no bairro de Perdizes, focando em imóveis com área de até 60m². Essa segmentação garante que a amostra seja relevante e comparável ao objeto de estudo, um apartamento de 56m². A localização comporta um raio de bairros adjacentes como parte componente, ainda que menos representativa, das proximidades.

A amostra bruta inicial, antes de qualquer pré-processamento, totalizou 449 registros. As estatísticas descritivas dessa amostra revelaram uma distribuição com uma alta dispersão e a presença de valores atípicos. Por exemplo, a média aritmética do preço por metro quadrado (R$/M2) era de $14.605,20, consideravelmente maior que a mediana de $13.703,70. O Coeficiente de Variação (CV) de 0,8127 (ou 81,27%) confirmou essa alta dispersão, indicando uma heterogeneidade significativa nos dados brutos. A distribuição por intervalos também evidenciava essa assimetria, com a maioria dos registros concentrada em um único intervalo de preços, enquanto alguns poucos anúncios se espalhavam por faixas de valor extremamente elevadas.

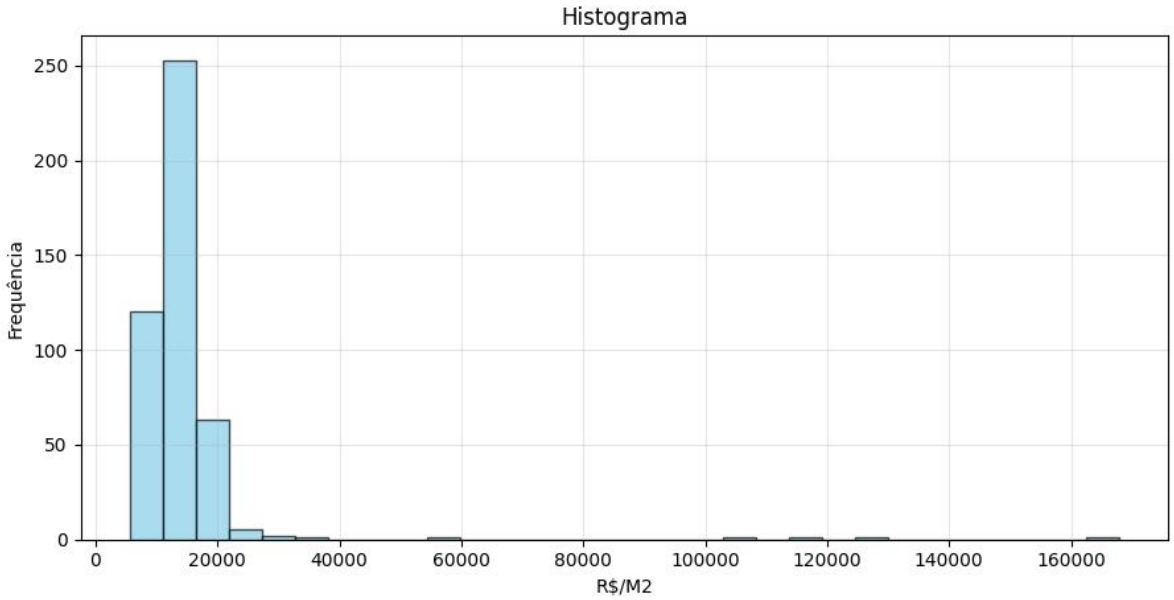

Figura 3. Histograma mostra a distribuição de dados de imóveis da população crua, sem tratamento.

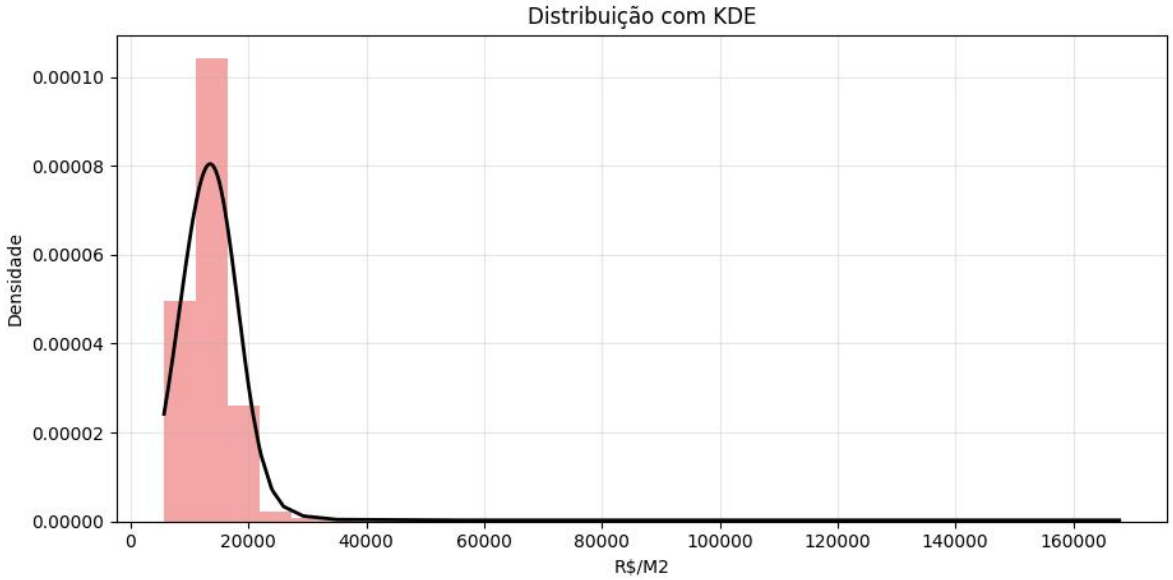

Figura 4. Gráfico de densidade probabilística revela alta concentração de valores em intervalos iniciais da população crua.

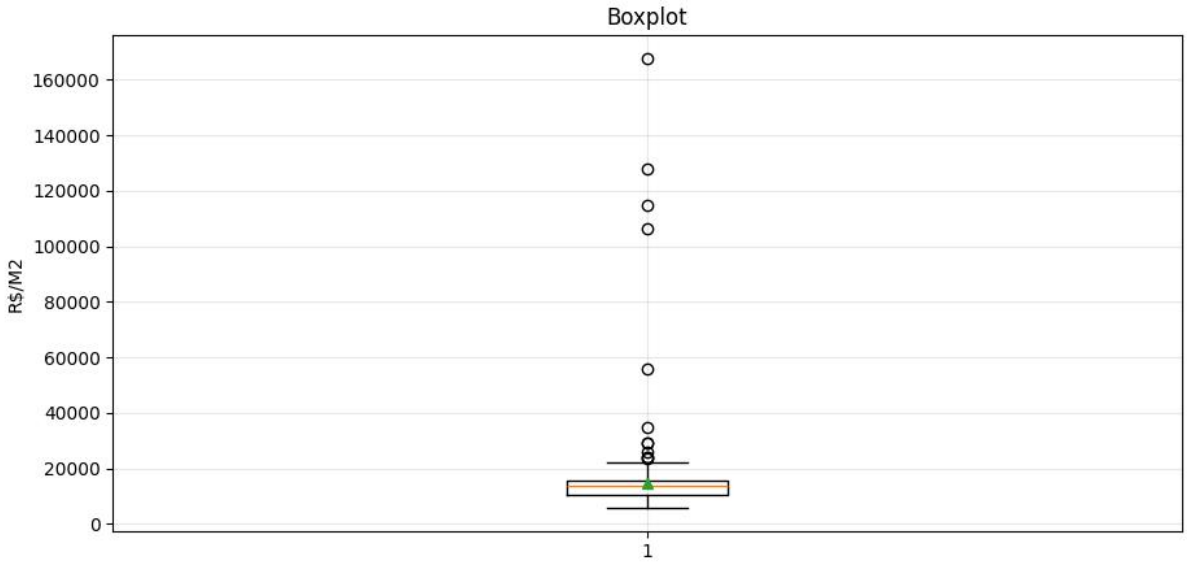

Figura 5. Boxplot mostra os valores considerados outliers da população crua.

O gráfico de Quantile-Quantile Plot na Figura 6 foi utilizado para avaliar se um conjunto de dados segue uma distribuição normal. Nele, os quantis teóricos da normalidade são comparados aos quantis observados nos dados. Caso a distribuição fosse aproximadamente normal, os pontos azuis tenderiam a se alinhar sobre a reta vermelha, que representa a referência ideal.

No entanto, observa-se que, embora a parte central dos dados acompanhe relativamente bem a linha, indicando que a maior parte das observações se distribui

de forma próxima à normalidade, as extremidades mostram um comportamento distinto. No lado esquerdo, o desvio é discreto, mas no lado direito há uma clara dispersão para cima, com pontos muito afastados da reta. Isso evidencia a presença de valores extremos (outliers) e uma assimetria positiva, caracterizada por uma cauda longa à direita.

Na prática, isso significa que, enquanto a maioria das observações concentra-se em valores moderados, existe um conjunto menor de registros que assume valores muito elevados, distorcendo a normalidade dos dados. Esse comportamento é típico de distribuições associadas a preços, rendas ou valores de mercado, em que poucos casos excepcionais alcançam números muito acima da média geral.

Portanto, os resultados do Q-Q Plot deixam claro que os dados analisados não seguem uma distribuição normal, sobretudo devido ao impacto dos valores extremos no limite superior.

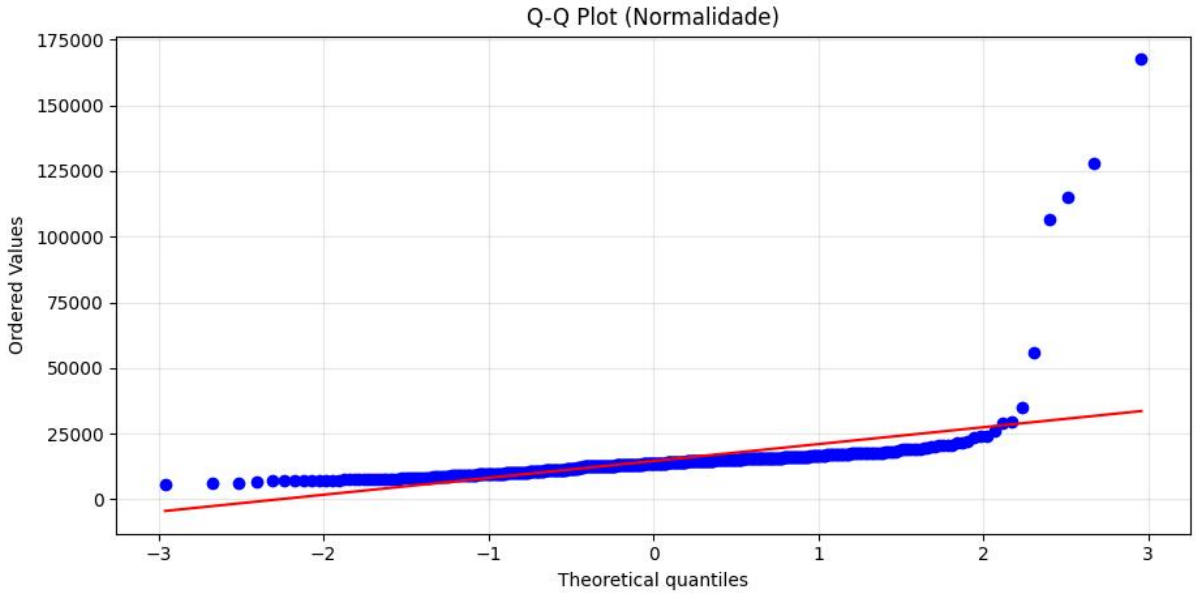

Figura 6. QQ-Plot atesta as observações de gráficos anteriores sobre a população crua não seguir uma distribuição normal

Para a obtenção de uma amostra final confiável e estatisticamente válida, foram aplicadas etapas de pré-processamento e limpeza de dados. O processo iniciou com a remoção de 65 registros duplicados, resultando em uma amostra de 384 imóveis. Em seguida, foi executado um processo iterativo para a substituição de outliers (valores discrepantes).

Iteração 1: O algoritmo de limpeza identificou que o Coeficiente de Variação (CV) da amostra, após a remoção de duplicatas, era de 0,8522. Por ser um valor alto, a mediana de $13.758,73 foi utilizada como valor de substituição. Nesse passo, 12 outliers, identificados pelo Método do Intervalo Interquartil (IQR), foram corrigidos. Após a substituição, o CV caiu para 0,2379, um indicativo de que a dispersão dos dados foi drasticamente reduzida.

Iteração 2: Com o CV já em um patamar baixo (0,2379), o sistema utilizou a média, que se tornou um estimador mais robusto. Nessa etapa, apenas 1 outlier foi identificado e substituído pelo valor da média, que era de $13.388,69. O CV se ajustou minimamente para 0,2359.

Iteração 3: O processo de limpeza foi repetido. O algoritmo não identificou mais outliers nem pelo método IQR nem pelo Z-Score. Com isso, foi atingida a convergência, e o processo de limpeza foi finalizado.

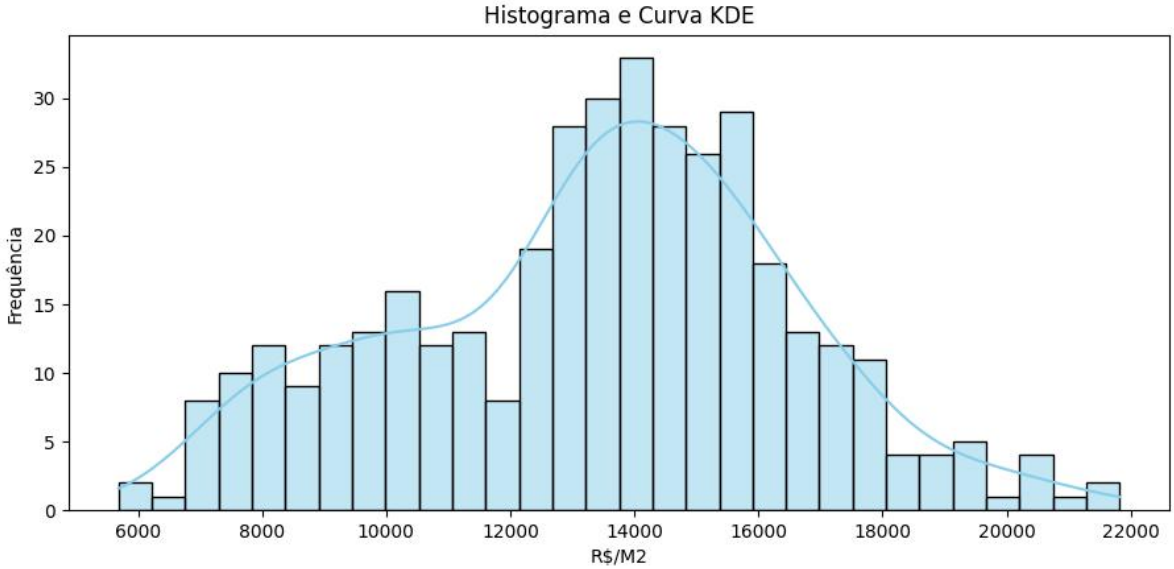

Figura 7. Gráfico de Histograma e Densidade mostra a distribuição de dados após o processamento e normalização.

A amostra final, limpa e padronizada, é composta por 384 registros. Após o pré-processamento, as estatísticas refletem uma distribuição muito mais homogênea e representativa do mercado. A média aritmética e a mediana do preço por metro quadrado (R$/M2) se aproximaram, atingindo $13.365,80 e $13.750,00, respectivamente. O Coeficiente de Variação final foi de 0,2359, confirmando que os dados agora possuem uma dispersão aceitável para a análise. A média ponderada por intervalos, um indicador mais preciso do preço típico da amostra, foi calculada

em $13.385,85. A distribuição de frequência dos dados, após a limpeza, mostra uma concentração significativa de imóveis nas faixas de preço mais representativas do mercado.

4.2 ANÁLISE EXPLORATÓRIA

Inicialmente, observa-se que as variáveis quantitativas essenciais — área (M2), preço (Preço) e o valor por metro quadrado (R$/M2) — estão completas e não contêm valores ausentes, formando uma base sólida para a análise estatística. Contudo, as informações de localização apresentam desafios significativos. A coluna Localidade é um texto não estruturado, e a coluna Endereço possui uma ausência de dados em 11,2% dos casos, limitando, por ora, análises que dependam de micro-segmentação geográfica precisa.

A amostra em questão é predominantemente composta por imóveis de pequeno a médio porte, com 50% das unidades possuindo até 54m² e 75% delas abaixo de 59m². Essa concentração em uma faixa de área específica é um traço definidor do perfil imobiliário analisado. O preço de venda, por sua vez, exibe a alta dispersão característica de ativos de mercado, com uma média de R$ 703.941,00 sendo elevada por imóveis de maior valor, um efeito evidenciado por uma mediana mais baixa, de R$ 650.000,00.

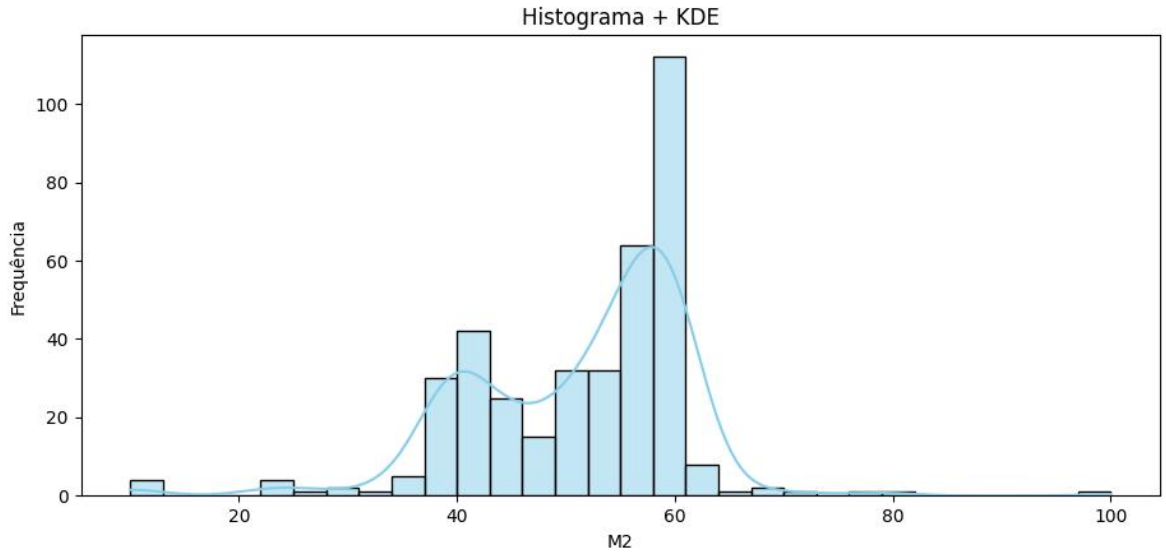

Figura 8. Gráfico de Histograma e Densidade exibe a distribuição de M2 na amostra; valores maiores que 60M2 servem de prova à inferência negativa de metragem como determinador de preço chamada de economia de escala (imóveis menores apresentam Preços por M2 mais elevados em relação a imóveis maiores por questão de liquidez)

Para uma comparação mais equitativa entre os imóveis, a métrica de valor por metro quadrado (R$/M2) se mostra mais estável e representativa. Com média e mediana muito próximas (R$ 13.365/m² e R$ 13.750/m², respectivamente), sua distribuição se aproxima da simetria. Uma análise mais aprofundada da distribuição deste indicador revela não apenas uma forte concentração de valores na faixa de R$ 12.800 a R$ 14.650, mas também uma notável consistência interna. De fato, a aplicação de testes estatísticos para detecção de outliers, como os métodos IQR e Z-Score, não identificou quaisquer pontos discrepantes que justificassem sua remoção ou alteração. Isso confere um grau elevado de confiança às medidas de tendência central como um retrato fiel do comportamento de preços neste segmento.

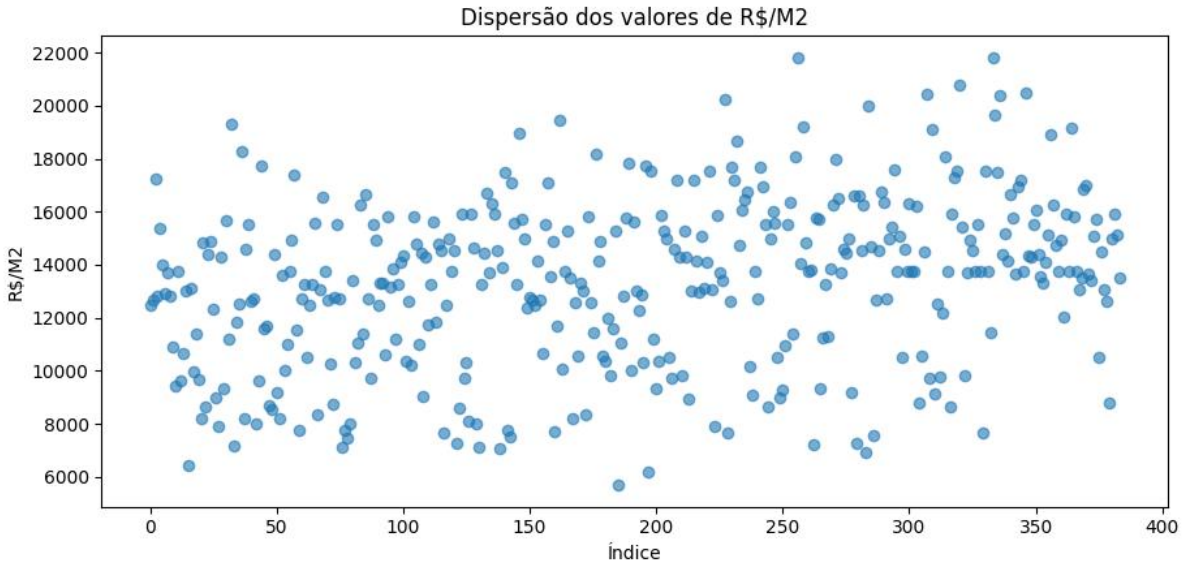

Figura 9. Gráfico de Dispersão exibe o espaço amostral dentro de uma distribuição próxima à normal

Com a consistência dos dados estabelecida, a análise avança para a relação entre as variáveis, onde se encontram os insights mais relevantes para a modelagem de preços. Contrariando uma intuição primária, a correlação entre a área de um imóvel (M2) e seu preço final (Preço) mostrou-se fraca, com um coeficiente de apenas 0.200. Este resultado demonstra que a área, isoladamente, possui um baixo poder preditivo sobre o valor de venda. A maior parte da variação nos preços é, portanto, explicada por outros fatores não capturados pelas variáveis numéricas disponíveis.

Complementar a isso, a correlação entre a área (M2) e o valor por metro quadrado (R$/M2) é fracamente negativa (-0.203). Este achado, embora sutil, é

consistente com o fenômeno de economia de escala frequentemente observado no mercado, onde unidades maiores tendem a possuir um custo marginal por metro quadrado ligeiramente inferior. A relação mais forte, como esperado, ocorre entre o preço e o valor por metro quadrado (0.656), uma vez que uma variável é derivada da outra.

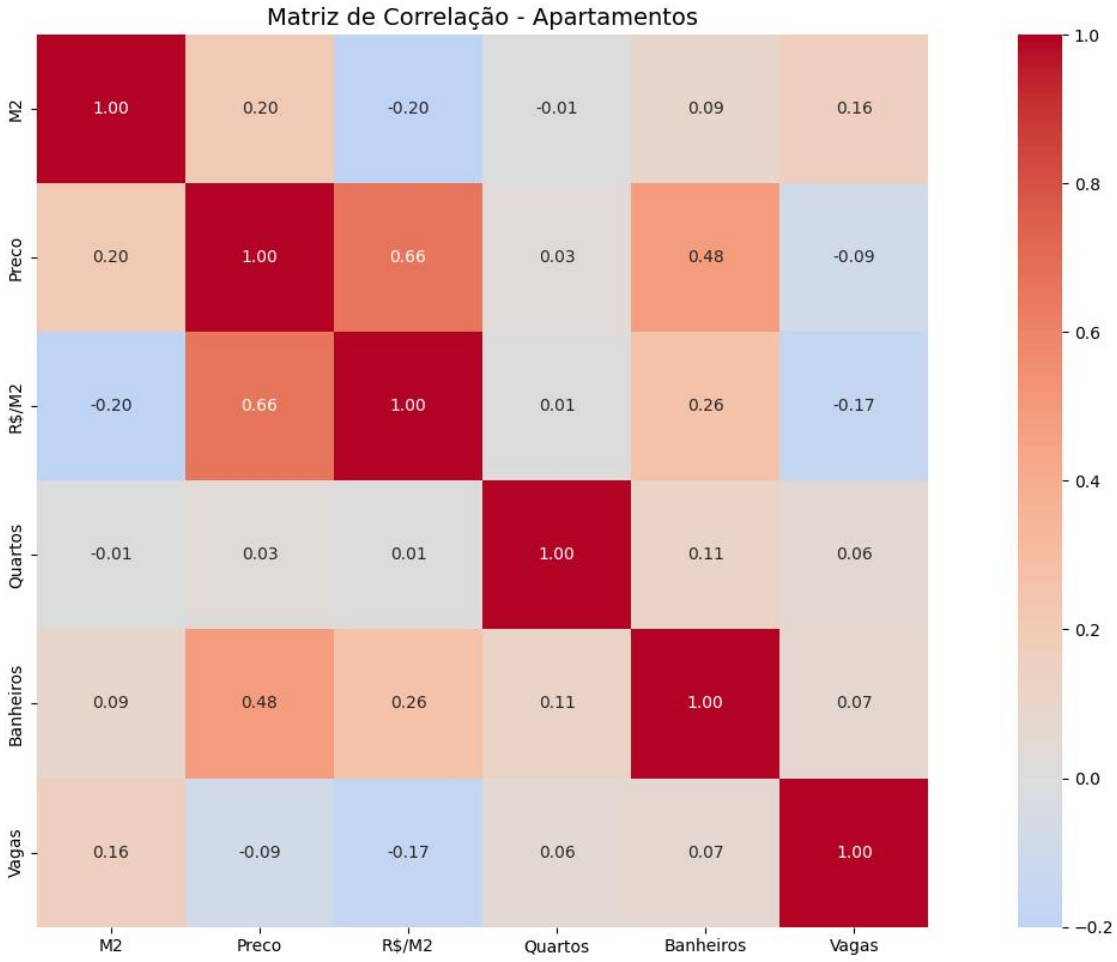

Figura 10. Matriz de Correlação entre as variáveis quantitativas do estudo

A análise por bairro revela uma heterogeneidade de valores que estava oculta nos dados agregados. A amostra é majoritariamente concentrada em Perdizes, que representa quase 60% dos imóveis, seguido por bairros como Água Branca, Sumarezinho e Vila Pompéia. Quando analisamos o valor médio por metro quadrado (R$/M2) em cada uma dessas regiões, as diferenças são notáveis:

Bairro	Nº de Imóveis	R$/M2 Médio	Preço Mediano
Sumarezinho	30	R$16.023	R$857.500
Perdizes	229	R$13.789	R$699.000
Vila Pompéia	21	R$12.924	R$600.000
Água Branca	65	R$11.258	R$515.000

Fica evidente que o Sumarezinho se destaca como a região de metro quadrado mais valorizado dentro da nossa amostra, superando a média de Perdizes em mais de 16%. Em contrapartida, Água Branca apresenta o valor médio mais baixo entre os bairros com representatividade, sendo em média 30% mais acessível que o Sumarezinho. Esta descoberta confirma com dados que a localização é, de fato, um fator preponderante na precificação, explicando grande parte da variação que a metragem (M2) sozinha não conseguia justificar.

A extração do número de quartos, banheiros e vagas também enriquece a análise. A grande maioria dos imóveis (85%) possui 2 quartos, caracterizando um perfil muito específico de apartamento. A influência dessas novas variáveis no preço pode ser observada na tabela abaixo, que utiliza a mediana para ser menos sensível a valores extremos:

Característica	Preço Mediano	R$/M2 Mediano
1 Quarto	R$525.000	R$14.857
2 Quartos	R$655.000	R$13.750
3 Quartos	R$1.150.000	R$12.916
<hr>		
1 Banheiro	R$625.000	R$13.805
2 Banheiros	R$785.000	R$13.529
<hr>		
Sem Vaga	R$625.000	R$14.166
1 Vaga	R$685.000	R$13.636
2 Vagas	R$1.349.500	R$13.125

A análise do dataset evidencia que o preço dos imóveis não está fortemente associado à metragem (M2) ou ao número de quartos, com correlações de apenas 0,20 e 0,03, respectivamente. Em contraste, o número de banheiros apresenta uma correlação moderada com o preço (0,48), indicando que imóveis com mais banheiros tendem a ter valores mais elevados. Isso sugere que, no mercado analisado, características funcionais e de conforto, como banheiros adicionais, podem ser mais valorizadas do que a simples metragem ou quantidade de quartos.

O valor por metro quadrado (R$/M2) apresenta uma correlação negativa com a metragem (-0,20), refletindo a tendência de imóveis menores terem um preço unitário mais alto. Essa observação é consistente com a chamada “economia de escala” no mercado imobiliário, onde unidades compactas, embora com preço total menor, apresentam maior valor unitário por área. A correlação positiva entre R$/M2 e preço (0,66) indica que imóveis mais caros também tendem a ter maior valor unitário, possivelmente devido à localização ou padrão construtivo.

Quanto às vagas de garagem, a correlação com o preço é ligeiramente negativa (-0,09), e sua relação com R$/M2 também é negativa (-0,17). Isso indica que, isoladamente, a presença de vagas não é um forte preditor de valorização linear no dataset analisado. No entanto, é possível que a influência das vagas esteja associada a outros fatores, como localização ou padrão do imóvel, que não são capturados diretamente pela correlação simples.

Observa-se que o número de quartos apresenta correlações muito baixas com todas as demais variáveis, sugerindo que a contagem de quartos, nesse conjunto específico, tem pouca relação linear com preço ou valor unitário.

Em síntese, a matriz de correlação demonstra que variáveis como número de banheiros e preço por metro quadrado têm associações mais consistentes com o preço do imóvel, enquanto métricas como número de quartos, metragem e vagas de garagem apresentam relações lineares fracas.

5. RESULTADOS

A aplicação do Método Comparativo de Mercado, apoiada em técnicas de ciência de dados, permitiu a obtenção de resultados estatisticamente robustos para a precificação do apartamento de 56m² em Perdizes. A amostra final, composta por 384 imóveis após o processo de limpeza e normalização, apresentou características estatísticas que conferem confiabilidade às estimativas realizadas.

5.1 Análise Descritiva da Amostra Final

O processo de pré-processamento demonstrou sua eficácia na obtenção de uma distribuição estatisticamente adequada. O Coeficiente de Variação final de 0,2359 (23,59%) representa uma redução significativa em relação aos 81,27% da amostra bruta, indicando uma dispersão controlada e representativa do mercado. A convergência entre média (R$ 13.365,80/m²) e mediana (R$ 13.750,00/m²) confirma a simetria da distribuição final, eliminando o viés de valores extremos que caracterizava os dados brutos.

A distribuição geográfica da amostra revelou heterogeneidade significativa entre os bairros analisados. Perdizes, representando 59,6% da amostra (229 imóveis), apresentou valor médio de R$ 13.789/m², posicionando-se como referência central para a análise. O Sumarezinho destacou-se com a maior valorização (R$ 16.023/m²), enquanto Água Branca apresentou o menor valor médio (R$ 11.258/m²), evidenciando diferenças de até 42% entre as micro-regiões estudadas.

5.2 Estimativa de Preço e Intervalo de Confiança

Para o apartamento objeto de estudo (56m² em Perdizes), a estimativa pontual baseada na média ponderada da amostra resulta em R$ 749.608,00. Este valor é derivado da aplicação da média ponderada por intervalos de R$ 13.385,85/m² sobre a área de 56m².

A aplicação do intervalo de confiança de 95% considera o desvio padrão de R$ 3.154,73/m² e o tamanho da amostra (n=384). Utilizando a distribuição t de Student para pequenas amostras, obtém-se o erro padrão da média de R$ 160,96/m². O valor crítico $t_{0,025}$ para 383 graus de liberdade é aproximadamente 1,966.

A margem de erro calculada é de R$ 316,45/m², resultando no seguinte intervalo de confiança para o valor por metro quadrado:

$IC_{95\%}$ = R$ 13.069,40/m² ≤ μ ≤ R$ 13.702,30/m²

Aplicando este intervalo ao imóvel de 56m², obtém-se:

$IC_{95\%}$ = R$ 731.886,40 ≤ Preço ≤ R$ 767.329,60

Portanto, com 95% de confiança estatística, o valor do apartamento de 56m² em Perdizes situa-se entre R$ 731.886,40 e R$ 767.329,60, com estimativa central de R$ 749.608,00.

5.3 Análise de Fatores Determinantes

A análise multivariada identificou que a localização constitui o principal fator determinante de preço, explicando grande parte da variação não capturada pela metragem. A correlação fraca entre área e preço (r=0,200) confirma que outros fatores exercem influência preponderante na formação de valor.

O número de banheiros emergiu como a característica física mais relevante (r=0,48 com o preço), superando significativamente a influência do número de quartos (r=0,03). Este achado sugere que elementos relacionados ao conforto e funcionalidade são mais valorizados pelo mercado do que a simples contagem de cômodos.

A verificação do fenômeno de economia de escala foi confirmada pela correlação negativa entre área e valor por metro quadrado (r=-0,203), indicando que imóveis menores tendem a apresentar maior valor unitário, possivelmente devido à maior liquidez no mercado.

5.4 Validação Estatística

Os testes de normalidade aplicados à amostra final confirmaram a adequação da distribuição para análises inferenciais. O Q-Q Plot da amostra tratada demonstrou alinhamento satisfatório com a distribuição normal teórica, validando o uso de intervalos de confiança baseados na distribuição t.

A ausência de outliers na amostra final, verificada pelos métodos IQR e Z-Score, garante que as medidas de tendência central representam adequadamente o comportamento típico do mercado, sem distorções causadas por casos extremos.

6. CONCLUSÃO

Este estudo demonstrou a viabilidade e eficácia da aplicação de métodos científicos de ciência de dados na precificação de ativos imobiliários, estabelecendo uma metodologia reprodutível e estatisticamente robusta para avaliação no mercado brasileiro. A estimativa obtida para o apartamento de 56m² em Perdizes, de R$ 749.608,00 ($IC_{95\%}$: R$ 731.886,40 - R$ 767.329,60), foi derivada de uma amostra representativa de 384 imóveis, processada através de algoritmos automatizados de coleta e tratamento de dados.

A comparação epistemológica entre as abordagens tradicionais de corretores e métodos científicos de ciência de dados, apresentada no texto preliminar, encontrou validação empírica nos resultados obtidos. A objetividade proporcionada pelo grande volume de dados (lei dos grandes números) e pelos critérios padronizados de seleção amostral resultou em estimativas com intervalo de confiança estatística bem definido, contrastando com a subjetividade inerente às avaliações baseadas em experiência pessoal e amostras reduzidas.

O processo de pré-processamento revelou-se fundamental para a qualidade dos resultados. A redução do Coeficiente de Variação de 81,27% para 23,59% através do tratamento iterativo de outliers permitiu a obtenção de uma distribuição adequada para análises inferenciais, eliminando distorções que poderiam comprometer a precisão das estimativas.

A identificação da localização como fator determinante primário na formação de preços corrobora conhecimentos estabelecidos do mercado imobiliário, enquanto a descoberta da baixa correlação entre metragem e preço (r=0,200) desafia intuições comuns sobre valorização imobiliária. A relevância do número de banheiros como preditor de preço (r=0,48) oferece insights práticos para avaliações futuras.

As diferenças significativas entre bairros adjacentes (até 42% entre Sumarezinho e Água Branca) evidenciam a importância da micro-segmentação geográfica em modelos de precificação, sugerindo que metodologias futuras devem incorporar variáveis de localização mais granulares.

A metodologia desenvolvida atende aos princípios fundamentais do método científico - reprodutibilidade, falseabilidade e ciclismo - extrapolando seu caráter corporativo para contribuir efetivamente com o conhecimento acadêmico do setor

imobiliário. O sistema automatizado de webscraping e o pipeline de tratamento ETL podem ser replicados em diferentes contextos geográficos e temporais, permitindo estudos comparativos e análises longitudinais.

As limitações identificadas, particularmente na estruturação de dados de localização (11,2% de valores ausentes) e na captura de variáveis qualitativas relevantes (padrão construtivo, estado de conservação, infraestrutura), apontam direções para desenvolvimento futuro. A incorporação de técnicas de processamento de linguagem natural para tratamento de descrições textuais e a integração de dados geoespaciais representam oportunidades de aprimoramento metodológico.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14724: informação e documentação – trabalhos acadêmicos – apresentação. Rio de Janeiro: ABNT, 2011.

BUSSAB, Wilton de Oliveira; MORETTIN, Pedro Alberto. Estatística básica. 9. ed. São Paulo: Saraiva, 2017.

COCHRAN, William Gemmell. Sampling techniques. 3. ed. New York: John Wiley & Sons, 1977.

DANTAS, Rubens Alves. Engenharia de avaliações: uma introdução à metodologia científica. 3. ed. São Paulo: PINI, 2012.

FIELD, Andy. Discovering statistics using IBM SPSS statistics. 5. ed. London: SAGE Publications, 2018.

GONZALEZ, Marco Aurélio Stumpf; FORMOSO, Carlos Torres. Mass appraisal with genetic fuzzy rule-based systems. Property Management, v. 24, n. 1, p. 20-30, 2006.

HAIR, Joseph F. et al. Multivariate data analysis. 8. ed. Andover: Cengage Learning, 2019.

KIMBALL, Ralph; ROSS, Margy. The data warehouse toolkit: the definitive guide to dimensional modeling. 3. ed. Indianapolis: John Wiley & Sons, 2013.

MITCHELL, Ryan. Web scraping with Python: collecting more data from the modern web. 2. ed. Sebastopol: O'Reilly Media, 2018.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. Introduction to linear regression analysis. 5. ed. Hoboken: John Wiley & Sons, 2012.

POPPER, Karl Raimund. The logic of scientific discovery. London: Routledge, 2013.

PROVOST, Foster; FAWCETT, Tom. Data science for business: what you need to know about data mining and data-analytic thinking. Sebastopol: O'Reilly Media, 2013.

TRIOLA, Mario F. Introdução à estatística: atualização da tecnologia. 12. ed. Rio de Janeiro: LTC, 2017.

ANEXO A – TÍTULO

AMOSTRA BRUTA____ Exemplo.cvs