UNIVERSIDADE FEDERAL DE SÃO CARLOS - CAMPUS SOROCABA
DEPARTAMENTO DE FÍSICA, QUÍMICA E MATEMÁTICA
LICENCIATURA EM MATEMÁTICA

Lucas Capovilla Cassimiro

UMA ANÁLISE EXPLORATÓRIA DE DADOS DO SETOR IMOBILIÁRIO

Sorocaba

2023

UNIVERSIDADE FEDERAL DE SÃO CARLOS - CAMPUS SOROCABA
DEPARTAMENTO DE FÍSICA, QUÍMICA E MATEMÁTICA
LICENCIATURA EM MATEMÁTICA

Lucas Capovilla Cassimiro

UMA ANÁLISE EXPLORATÓRIA DE DADOS DO SETOR IMOBILIÁRIO

Trabalho de Conclusão de Curso de Graduação apresentado como requisito parcial para a obtenção do título de licenciado em Matemática sob a orientação do Prof. Dr. Antonio Luis Venezuela.

Sorocaba

2023

Capovilla, Lucas

Uma análise exploratória de dados do setor imobiliário / Lucas Capovilla -- 2023.
42f.

TCC (Graduação) - Universidade Federal de São Carlos, campus Sorocaba, Sorocaba
Orientador (a): Antonio Luis Venezuela
Banca Examinadora: Renato Fernandes Cantão, Magda da Silva Peixoto
Bibliografia

1. Correlação de Pearson. 2. Python. 3. Estatística. I. Capovilla, Lucas. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Maria Aparecida de Lourdes Mariano - CRB/8 6979

FUNDAÇÃO UNIVERSIDADE FEDERAL DE SÃO CARLOS

COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA DE SOROCABA - CCML-So/CCTS Rod. João Leme dos Santos km 110 - SP-264, s/n - Bairro Itinga, Sorocaba/SP, CEP 18052-780 Telefone: (15) 32298874 - http://www.ufscar.br

DP-TCC-FA nº 8/2023/CCML-So/CCTS

Graduação: Defesa Pública de Trabalho de Conclusão de Curso

Folha Aprovação (GDP-TCC-FA)

FOLHA DE APROVAÇÃO

LUCAS CAPOVILLA CASSIMIRO

UMA ANÁLISE EXPLORATÓRIA DE DADOS DO SETOR IMOBILIÁRIO

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos – Campus Sorocaba

Sorocaba, 05 de setembro de 2023

ASSINATURAS E CIÊNCIAS

Cargo/Função	**Nome Completo**
Orientador	Prof. Dr. Antonio Luís Venezuela
Membro da Banca 1	Profa. Dra. Magda da Silva Peixoto
Membro da Banca 2	Prof. Dr. Renato Fernandes Cantão

Documento assinado eletronicamente por **Antonio Luis Venezuela**, **Docente**, em 06/09/2023, às 13:01, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Antonio Luis Venezuela**, **Docente**, em 06/09/2023, às 13:01, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Antonio Luis Venezuela**, **Docente**, em 06/09/2023, às 13:01, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufscar.br/autenticacao, informando o código verificador **1169611** e o código CRC **8FC84B72**.

Referência: Caso responda a este documento, indicar expressamente o Processo nº 23112.008909/2021-87 SEI nº 1169611

Modelo de Documento: Grad: Defesa TCC: Folha Aprovação, versão de 02/Agosto/2019

AGRADECIMENTOS

Inicialmente, expresso minha gratidão a Deus por ter me guiado até este momento e por todas as dificuldades que foram superadas. Também quero agradecer meus pais, Claudia, Eduardo e minha irmã, Ruth, por todo apoio, parceria e aprendizados que tornaram possível a realização desse objetivo de me formar em uma Universidade Pública.

Sou grato à minha parceira, Laís, por toda a paciência, atenção e apoio fornecendo todo o suporte necessário em diversos momentos. Também quero agradecer a cada amigo que participou dessa jornada.

Não posso deixar de agradecer a cada professor que participou da minha formação ao longo da graduação, ensinando muito além de uma grade curricular e enriquecendo a vida de cada aluno. Em específico, gostaria de agradecer meu orientador, Prof. Dr. Antonio Luis Venezuela, que me deu todo o apoio necessário desde o início da realização deste trabalho.

Gostaria de estender meu agradecimento à Universidade Federal de São Carlos (UFSCar), campus Sorocaba, bem como cada funcionário que permite a boa permanência e o ensino de qualidade, em particular à Rafaela Marie, secretária do curso de Licenciatura em Matemática que esteve sempre pronta a ajudar e orientar cada aluno.

RESUMO

Este trabalho desenvolvido com uma base de dados disponibilizada pelo website voltado à análise de dados chamado Kaggle, tem como objetivo desenvolver uma análise exploratória através de ferramentas estatísticas para responder a perguntas de negócio que impactam diretamente no dia a dia da empresa fictícia House Rocket. O banco de dados é composto por 21612 entradas que são imóveis e suas informações complementares como metragem quadrada, data da última reforma, valor de venda, entre outras. Para desenvolver este trabalho foram sondados outros trabalhos realizados previamente, mas que também usaram bases de dados disponibilizadas pelo Kaggle ou de análise de dados para trazer valor a um negócio ou embasar decisões. Por fim, constatou-se que utilizar conhecimentos estatísticos aliados à tecnologias digitais pode ser uma ferramenta poderosa para a solução dos mais variados problemas.

Palavras-chave: Kaggle; Estatística; Imóveis; Análise Exploratória.

ABSTRACT

This work, developed with a database through the data analysis website called Kaggle, aims to perform an exploratory analysis using statistical tools to answer business questions that directly impact the day-to-day of the fictitious company House Rocket. The database is composed of 21,612 entries which are properties and their complementary information such as square footage, last renovation date, selling price, among others. To develop this work, other previously conducted works were surveyed, which also used databases provided by Kaggle or finger analysis to bring value to a business or support decisions. Finally, it was found that using statistical knowledge combined with digital technologies can be a powerful tool for solving various problems.

Key words: Kaggle; Statistic, Properties; Exploratory Analysis.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A análise de dados é um processo que busca extrair informações relevantes a um determinado negócio ou objetivo a partir de um determinado volume de dados. Com a evolução da tecnologia e a geração cada vez maior de informações vindas de diferentes fontes, é de extrema importância o uso da análise de dados com o objetivo de acelerar a tomada de decisão e trazer vantagem competitiva para as empresas. Consequentemente, o mercado de trabalho busca profissionais que estejam aptos a traduzir tantas informações em conhecimento e, por isso, senti-me motivado a estudar mais sobre tal assunto.

Pode-se classificar a análise de dados em quatro tipos, sendo elas preditiva, prescritiva, descritiva ou diagnóstica. A análise preditiva, como o próprio nome diz, tem o objetivo de realizar projeções de cenários, identificando padrões ou tendências que possam ser importantes para um determinado fim. Já a prescritiva tem como objetivo avaliar quais as consequências que as decisões podem causar, buscando assim a melhor performance para atingir um objetivo específico. O foco de uma análise descritiva é descrever eventos ou objetos através desse estudo, já a análise diagnóstica busca resolver um problema ou encontrar a causa de algum efeito analisado.

É importante observar que uma boa análise começa com um sólido processo de *ETL (Extract Transform Load)* que tem por objetivo fazer todo o tratamento necessário para que a base de dados esteja pronta e organizada da melhor forma para que as análises sejam realizadas.

Neste trabalho foi desenvolvida uma análise estatística exploratória dos dados disponíveis no site Kaggle, no banco de dados nomeado como *House Sales in King County*[1]. Os dados disponibilizados pelo Kaggle sempre trazem as informações randomizadas, pois muitas vezes foram disponibilizados por empresas reais com informações do negócio e às vezes até de clientes. Será adotado então como *House Rocket* o nome desta empresa fictícia posicionada no setor imobiliário que está constantemente buscando boas oportunidades de negócio na compra e venda de imóveis (KAGGLE, 2023).

Com o objetivo de agregar valor ao negócio da *House Rocket*, essa análise além de compreender o comportamento do mercado imobiliário no intervalo de tempo que os dados apresentam, visa responder a três perguntas de negócio, sendo elas:

1) Quais recursos de um imóvel mais impactam no preço?
2) A House Rocket poderia fazer uma reforma para aumentar o preço da venda? Quais seriam as sugestões de mudanças?

Conforme veremos abaixo, alguns autores já produziram conteúdos semelhantes sobre análise de dados, regressão linear ou logística, utilizando um banco de dados aberto. Esses conteúdos serviram de referência e ponto de partida para o desenvolvimento deste trabalho por discutirem assuntos semelhantes e pertinentes ao tema desenvolvido.

O primeiro trabalho foi sobre a análise das características de jogabilidade no Player Unknown's Battleground (PUBG) que é um jogo online usando a árvore de decisão, obra do autor Santos Júnior (2019), nele o autor analisa as características de jogabilidade e identificar quais fatores influenciam mais na vitória dos jogadores. Para realizar a análise, foi utilizada uma árvore de decisão como modelo de classificação. Os dados foram coletados a partir de partidas do jogo registradas no servidor oficial do *PUBG*. Foram analisados um total de 27 atributos, incluindo pontuação, tempo de sobrevivência, quantidade de mortes, quantidade de jogadores eliminados, entre outros.

Os resultados da análise da árvore de decisão indicaram que o atributo mais importante para determinar a vitória no PUBG é a pontuação, seguido pelo tempo de sobrevivência e a quantidade de jogadores eliminados. Outros atributos, como quantidade de mortes e distância percorrida tiveram menos influência na vitória. Os autores concluem que a análise das características de jogabilidade pode ser útil para entender como os jogadores interagem com o jogo e identificar estratégias que levem à vitória. Além disso, a utilização de uma árvore de decisão como modelo de classificação pode ser uma ferramenta eficaz para a análise de dados de jogos (Santos Júnior, 2019).

O segundo trabalho foi de Marra (2019) sobre previsão de dificuldades financeiras em empresas latino-americanas. O foco foi criar um modelo que previsse a incidência desses desafios. Usando um banco de dados repleto de informações financeiras e contábeis dessas corporações, o estudo procurou identificar quais fatores têm maior influência através de métricas, como índices de liquidez, rentabilidade, endividamento e atividade, foram consideradas. Posteriormente, o modelo conhecido como 'random forest' foi empregado (Marra, 2019).

O 'random forest' revelou uma acurácia de 80% na previsão de corporações enfrentando desafios financeiros. Os pesquisadores concluíram que o uso de modelos de aprendizado de máquina pode ser uma ferramenta valiosa para antecipar problemas financeiros em corporações latino-americanas (Marra, 2019).

O terceiro trabalho foi de Vaz (2016), que foca no emprego de regressão linear em análise de dados, detalhando o procedimento de criação e avaliação de modelos baseados nessa abordagem. O estudo inicia com a definição de regressão linear e seus vários tipos. Posteriormente, o autor explora a montagem de modelos de regressão, ressaltando as fases do processo que engloba a seleção de variáveis, a avaliação da qualidade dos dados e a validação do modelo. Além disso, Vaz descreve diferentes técnicas de avaliação para modelos de regressão linear, como a validação cruzada, o método bootstrap e critérios de seleção de modelos.

Os prós e contras de cada método de avaliação de modelos de regressão linear são discutidos. Por fim, o trabalho conclui que a regressão linear é uma técnica poderosa em análise de dados, mas que a qualidade dos resultados depende da qualidade dos dados e da escolha adequada de variáveis. O trabalho recomenda a utilização de métodos de validação para garantir que o modelo construído seja válido e útil para a tomada de decisões (Vaz, 2016).

O quarto trabalho foi de Freitas (2019) e começa com uma revisão da literatura sobre a regressão logística, apresentando os conceitos básicos e suas aplicações em diferentes áreas. O estudo descreve a metodologia utilizada para coletar e pré-processar os dados, incluindo a seleção de variáveis e a normalização. Esse artigo apresenta então o processo de construção do modelo de regressão logística, incluindo a seleção de variáveis, a definição da função de custo e a

escolha do algoritmo de otimização. O modelo foi avaliado por meio de métricas como acurácia, precisão, recall e F1-score. Os resultados do estudo indicam que o modelo de regressão logística teve um bom desempenho na modelagem do risco de diabetes em mulheres, alcançando uma acurácia de mais de 70%. E que é uma técnica poderosa em análise de dados, especialmente para problemas de classificação binária como o estudo de caso apresentado (Freitas, 2019).

Neste trabalho segue-se a seguinte estrutura: no capítulo 2 são abordados os principais conceitos necessários para a realização da análise apresentada no trabalho, acompanhados de exemplos práticos em algumas situações. No capítulo 3 é apresentada a metodologia para desenvolver a análise, comentando sobre todas as informações presentes na base de dados bem como os passos para responder às perguntas de negócio. No capítulo 4 encontra-se o centro deste trabalho, onde são apresentados os resultados após uma exploração do banco de dados e seus devidos desdobramentos para responder às perguntas apresentadas anteriormente.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão discutidos conhecimentos necessários para desenvolver a análise dos dados disponibilizados, sejam eles uma interpretação gráfica, uma definição teórica ou uma habilidade prática de programação.

2.1 CORRELAÇÃO E CAUSALIDADE

Esta seção de fundamentação teórica tem como objetivo fornecer uma base sólida sobre correlação e causalidade, elucidando suas definições, diferenças e principais desafios na análise empírica (Morettin e Bussab, 2010).

No campo da estatística é importante compreender os diferentes tipos de variáveis que descrevem as características dos dados coletados. Existem dois principais tipos de variáveis estatísticas: as qualitativas ou categóricas e as quantitativas.

As variáveis qualitativas, também conhecidas como categóricas, representam características ou qualidades distintas e não podem ser mensuradas numericamente. Dentro dessa categoria, podemos identificar duas subcategorias. A primeira delas é composta pelas variáveis nominais, que não possuem uma ordem específica. Exemplos comuns são o estado civil (solteiro, casado, divorciado) ou a cor dos olhos (azul, verde, castanho). A segunda subcategoria é composta pelas variáveis ordinais, que possuem uma ordem específica. Exemplos incluem a classificação de satisfação do cliente (muito satisfeito, satisfeito, insatisfeito) ou a escala de dor (nenhuma dor, leve, moderada, intensa).

Por outro lado, as variáveis quantitativas representam quantidades ou medidas numéricas. Essas variáveis podem ser subdivididas em duas categorias. A primeira delas é composta pelas variáveis discretas, que assumem valores inteiros e não podem ser subdivididas em valores menores. Exemplos comuns incluem o número de filhos em uma família ou o número de acidentes de trânsito em uma cidade durante um determinado período de tempo. A segunda categoria é composta

pelas variáveis contínuas, que assumem valores em um intervalo contínuo e podem ser subdivididas em valores menores. Exemplos incluem a altura de uma pessoa, o tempo gasto em uma tarefa ou a temperatura de um ambiente. A base de dados utilizada é composta por variáveis quantitativas conforme a Tabela 1.

Tabela 1: Lista de variáveis.

Variáveis
id
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade
sqft_above
sqft_basement
yr_built
yr_renovated
zipcode
lat
long
sqft_living15
sqft_lot15
month
year

Fonte: Autoria própria.

A relação entre correlação e causalidade é um tema crucial na pesquisa científica, especialmente quando se busca compreender a natureza dos fenômenos e estabelecer relações de causa e efeito.

A correlação linear é uma medida estatística que descreve a relação entre duas variáveis, indicando o grau de associação linear entre elas. É comumente representada pelo coeficiente de correlação de Pearson, que varia de -1 a 1. Um valor próximo de 1 indica uma correlação positiva forte, enquanto um valor próximo

de -1 indica uma correlação negativa forte. Por outro lado, um valor próximo de zero sugere uma correlação fraca ou inexistente entre as variáveis. Por sua vez, a causalidade refere-se à relação de causa e efeito entre duas variáveis, ou seja, uma variável influencia diretamente a outra. Estabelecer causalidade é um processo complexo, pois requer a identificação de uma conexão causal plausível, além de evidências empíricas consistentes.

Embora a correlação seja uma ferramenta útil para explorar associações entre variáveis, ela não é suficiente para inferir causalidade, uma vez que a presença de correlação não implica necessariamente uma relação causal. Existem diferenças fundamentais entre correlação e causalidade. A correlação descreve a força e direção da associação entre duas variáveis, enquanto a causalidade envolve uma relação de causa e efeito, em que uma variável é considerada como influenciadora da outra. A inferência de correlação pode ser identificada por meio de técnicas estatísticas, como o coeficiente de correlação, enquanto a inferência de causalidade requer uma análise mais rigorosa, envolvendo o controle de variáveis de confusão, experimentos controlados e a aplicação de métodos como o design experimental ou análise de regressão causal.

No entanto, há desafios na inferência de causalidade. Um deles é a possibilidade de relações espúrias, em que outras variáveis não observadas podem estar influenciando tanto as variáveis em questão, levando a uma associação aparente. Além disso, existem casos em que duas variáveis podem estar correlacionadas indiretamente, por meio de uma terceira variável que atua como mediadora. Nesses casos, a correlação entre as variáveis diretas pode ser confundida como uma relação causal, quando na verdade é mediada por uma terceira variável. Estabelecer uma relação causal requer evidências mais robustas, como experimentos controlados, randomização e manipulação das variáveis independentes, além de considerações teóricas e conhecimento especializado.

Na Tabela 2 serão apresentados todos os casos referentes aos graus de correlação.

Tabela 2: Graus de correlação.

Coeficiente de Correlação Linear			
Negativa		**Positiva**	
$-0,95 < r \leq -1$	Correlação negativa muito forte	$0,95 < r \leq 1$	Correlação positiva muito forte
$-0,50 < r \leq -0,95$	Correlação negativa forte	$0,50 < r \leq 0,95$	Correlação positiva forte
$-0,10 < r \leq -0,50$	Correlação negativa moderada	$0,10 < r \leq 0,50$	Correlação positiva moderada
$0 < r \leq -0,10$	Correlação negativa fraca	$0 < r \leq 0,10$	Correlação positiva fraca

Fonte: Autoria própria.

Neste trabalho usaremos a correlação linear de Pearson para desenvolver a análise, sendo N a quantidade de valores, (1) coeficiente de correlação linear de Pearson, (2) escore z de um sujeito particular na variável X, (3) escore z de um sujeito particular na variável Y, (4) desvio padrão de X, (5) desvio padrão de Y, (6) média de valores de X e (7) média de valores de Y:

$$r = \frac{1}{N}\sum_{i=1}^{N} z_{x_i} z_{y_i} \tag{1}$$

$$z_{x_i} = \frac{x_i - \bar{x}}{S_x} \tag{2}$$

$$z_{y_i} = \frac{y_i - \bar{y}}{S_y} \tag{3}$$

$$S_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x^2 - \bar{x}^2} \tag{4}$$

$$S_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N} y^2 - \bar{y}^2} \tag{5}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (7)$$

Para facilitar a compreensão, abaixo será aplicado um exemplo de como calcular a correlação de Pearson, de acordo com Bussab e Morettin (2010). Na Tabela 3 temos o número de anos escolares cursados pelo pai (X) e o número de anos escolares cursados pelo filho (Y).

Tabela 3: Anos escolares pais e filhos.

Criança	x_i	y_i
A	12	12
B	10	8
C	6	6
D	16	11
E	8	10
F	9	8
G	12	11
Soma	73	66

Fonte: Autoria própria.

Para aplicar a fórmula da correlação de Pearson é necessário primeiro calcular as seguintes estatísticas: $\overline{X}$, $\overline{Y}$, z_x e z_y.

Tabela 4: Anos escolares pais e filhos com as estatísticas x_i^2, y_i^2 e $x_i y_i$ calculadas.

i	Criança	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	A	12	12	144	144	144
2	B	10	8	100	64	80
3	C	6	6	36	36	36
4	D	16	11	256	121	176
5	E	8	10	64	100	80
6	F	9	8	81	64	72
N = 7	G	12	11	144	121	132
	Total	**73**	**66**	**825**	**650**	**720**

Fonte: Autoria própria.

$$\overline{x} = 10,43$$

$\overline{y} = 9,43$

$S_x = 3,01$

$S_y = 1,98$

Tabela 5: Anos escolares pais e filhos com as estatísticas $\overline{X}$, $\overline{Y}$, z_x e z_y calculadas.

Criança	X	X - X'	(X - X')/Sx	Y	Y - Y'	(Y - Y')/SY	ZxZy
A	12	1,57	0,52	12	2,57	1,29	0,67
B	10	-0,43	-0,14	8	-1,43	-0,72	0,10
C	6	-4,43	-1,47	6	-3,43	-1,72	2,53
D	16	5,57	1,85	11	1,57	0,79	1,46
E	8	-2,43	-0,80	10	0,57	0,29	-0,23
F	9	-1,43	-0,47	8	-1,43	-0,72	0,34
G	12	1,57	0,52	11	1,57	0,79	0,41
Soma	73			66			5,28

Fonte: Autoria própria.

$r = 0,75$

De acordo com a tabela de dados o coeficiente de Pearson ser 0,75 significa uma correlação razoavelmente forte entre o nível educacional que os pais atingem e o nível educacional que os filhos atingem. Isso significa que olhando para cada caso especificamente, podemos observar que se os pais atingiram um grau escolar maior, há uma tendência de os filhos atingirem um grau maior de escolaridade também.

2.3 KAGGLE

Nesta seção utilizaremos como referência KAGGLE (2023).

O Kaggle é uma plataforma online fundada em 2010 por Anthony Goldbloom e Ben Hamner. Seu objetivo principal é fornecer uma comunidade para cientistas de dados. A plataforma oferece uma ampla variedade de conjuntos de dados e competições de aprendizado de máquina. Os usuários têm acesso a recursos como desafios, fóruns de discussão, podem criar perfis, competir em desafios e colaborar

com outros membros além de facilitar a conexão entre usuários e empresas em busca de talentos em ciência de dados.

Os conjuntos de dados disponíveis abrangem diversas áreas, como ciência, tecnologia, saúde e finanças, variando em tamanho e complexidade. Além disso, a plataforma oferece competições de aprendizado de máquina patrocinadas por empresas, nas quais os participantes desenvolvem modelos para resolver problemas específicos.Os participantes têm prazos para criar e submeter seus modelos, que são avaliados com base em precisão e eficácia. Os vencedores das competições recebem prêmios e reconhecimento. Além dos conjuntos de dados e competições, o Kaggle oferece recursos de aprendizado, como tutoriais, cursos e palestras.

2.3 PYTHON

Nesta seção utilizaremos como referência PYTHON (2023).

O Python é uma linguagem de programação de alto nível e amplamente utilizada em diversos domínios, como ciência de dados, desenvolvimento web, automação e robótica. Ele é conhecido por sua facilidade de uso, clareza de sintaxe e suporte da comunidade. O Python é multiplataforma, tem uma vasta biblioteca padrão e uma comunidade ativa de usuários e desenvolvedores.

Na área de ciência de dados, o Python é usado com bibliotecas como o *Pandas (2023)*, que oferece recursos para análise e manipulação de dados, incluindo estruturas de dados flexíveis, leitura e escrita de arquivos de dados e manipulação de séries temporais. A biblioteca Matplotlib *(2023)* é amplamente utilizada para visualização de dados, com uma variedade de gráficos disponíveis e recursos de personalização. A biblioteca Seaborn *(2023)* é baseada no *Matplotlib (2023)* e fornece uma interface de alto nível para criar gráficos estatísticos atraentes e informativos.

Em resumo, o Python é uma linguagem de programação versátil e poderosa, com uma variedade de bibliotecas disponíveis para análise, manipulação e

visualização de dados. Sua comunidade ativa e sua sintaxe clara tornam o Python uma escolha popular para profissionais em várias áreas.

3 METODOLOGIA

Aqui serão abordados em detalhes os elementos metodológicos empregados no desenvolvimento da pesquisa. Serão discutidas as abordagens, técnicas de coleta de dados, análise e interpretação utilizadas para atender aos objetivos propostos.

3.1 BANCO DE DADOS

Para a realização deste trabalho, os dados foram coletados do Kaggle, cujo endereço está presente nas referências bibliográficas. O banco de dados coletado é composto por 21 variáveis compondo dados de diferentes imóveis comercializados entre 2014 e 2015. Inicialmente o banco possui 21612 cadastros, porém foram feitas algumas verificações e tratamentos de dados reduzindo algumas linhas totalizando 20012 cadastros (LUCASCAPOVILLA, 2023).

3.2 VARIÁVEIS ANALISADAS

Foram consideradas para a análise 21 variáveis relacionadas aos imóveis negociados. Essas variáveis representam desde elementos gerais dos imóveis como número de banheiros e metragem quadrada, além de valores, até informações um pouco mais aprofundadas como vista, condição da construção e nota. Na Tabela 6 estão detalhadas todas as variáveis do banco de dados e suas respectivas descrições.

Tabela 6: Descrição da base de dados.

Campo	Descrição
id	ID único para cada casa vendida
date	data em que a casa foi vendida
price	preço de venda da casa
bedrooms	número de quartos
bathrooms	número de banheiros, onde 0,5 conta como um cômodo com toalete, porém sem chuveiro
sqft_living	metragem quadrada da área interna da construção
sqft_lot	metragem quadrada do terreno
floors	número de andares
waterfront	variável para saber se a casa tem vista para o mar
view	uma nota de 0 a 4 sobre quão boa é a vista da construção
condition	uma nota de 1 a 5 sobre a condição da construção
grade	uma nota de 1 a 13, onde 1-3 representa uma nota de construção e design baixa, 7 é o valor mediano e 11-13 uma construção de alto nível e design
sqft_above	metragem quadrada do interior da casa dos andares acima do térreo
sqft_basement	metragem quadrada do interior da casa dos andares abaixo do térreo
yr_built	ano de construção da casa
yr_renovated	ano da última reforma
zipcode	o código postal
lat	latitude
long	longitude
sqft_living15	metragem quadrada média do interior da construção dos 15 vizinhos mais próximos
sqft_lot15	metragem quadrada do terreno dos 15 vizinhos mais próximos

Fonte: Autoria própria.

3.3 OUTLIERS

Como pode ser observado nas Figuras 2, 3 e 4 existem alguns outliers no banco de dados, que são pontos que se destacam dos demais apresentando valores discrepantes sendo muito maiores ou muito menores do que a maior parte dos valores. Para entender quais impactos esses valores apresentam na análise são realizadas algumas comparações através das informações obtidas antes e após a remoção dos valores discrepantes.

3.4 DESENVOLVIMENTO DA ANÁLISE

Para realizar o tratamento dos dados e análise exploratória foi utilizada a linguagem de programação Python. Foram geradas algumas visualizações gráficas como gráfico de barras e boxplot para compreender o universo de valores dos imóveis, em seguida foi utilizada a Correlação de Pearson, que permitiu identificar as variáveis mais relevantes em relação ao preço, ou seja, quais variáveis apresentam maior correlação positiva ou negativa.

Posteriormente, foi analisada a possibilidade de realizar reformas para aumentar o preço de venda através da correlação de Pearson, tendo em mente que o objetivo final sempre é o lucro com a negociação dos imóveis. É importante entender se uma vez adquirido o imóvel, vale a pena realizar alguma reforma e em caso positivo, qual reforma deve ser feita.

4 RESULTADOS E ANÁLISE

Neste capítulo através da exploração de dados utilizando ferramentas como a correlação de Pearson e visualizações gráficas serão discutidas as três perguntas de negócio a seguir:

1) Quais recursos de um imóvel mais impactam no preço?
2) A House Rocket poderia fazer uma reforma para aumentar o preço da venda? Quais seriam as sugestões de mudanças?

Para começar, foi produzida uma tabela que traz um resumo dos tipos de dados de cada variável e a contagem de campos nulos e, em seguida, uma segunda tabela com algumas medidas estatísticas de cada campo como contagem, média, desvio padrão, valores mínimo e máximo e os quartis, conforme as Tabelas 6 e 7.

Tabela 7: Tabela descritiva do banco de dados com algumas medidas estatísticas pré tratamento de outliers.

Variáveis	média	desv. padrão	mínimo	25%	50%	75%	máximo
id	-	-	-	-	-	-	-
price	540088.14	367127.20	75000.00	321950.00	450000.00	645000.00	7700000.00
bedrooms	3.37	0.93	0.00	3.00	3.00	4.00	33.00
bathrooms	02.11	0.77	0.00	1.75	2.25	2.50	8.00
sqft_living	2079.90	918.44	290.00	1427.00	1910.00	2550.00	13540.00
sqft_lot	15106.97	41420.51	520.00	5040.00	7618.00	10688.00	1651359.00
floors	1.49	0.54	1.00	1.00	1.50	2.00	3.50
waterfront	0.01	0.09	0.00	0.00	0.00	0.00	1.00
view	0.23	0.77	0.00	0.00	0.00	0.00	4.00
condition	3.41	0.65	1.00	3.00	3.00	4.00	5.00
grade	7.66	1.18	1.00	7.00	7.00	8.00	13.00
sqft_above	1788.39	828.09	290.00	1190.00	1560.00	2210.00	9410.00
sqft_basement	291.51	442.58	0.00	0.00	0.00	560.00	4820.00
yr_built	25934	29.37	1900.00	1951.00	1975.00	1997.00	2015.00
yr_renovated	84.40	401.68	0.00	0.00	0.00	0.00	2015.00
zipcode	98077.94	53.51	98001.00	98033.00	98065.00	98118.00	98199.00
lat	47.56	0.14	47.16	47.47	47.57	47.68	47.78
long	-122.21	0.14	-122.52	-122.33	-122.23	-122.12	-121.31
sqft_living15	1986.55	685.39	399.00	1490.00	1840.00	2360.00	6210.00
sqft_lot15	12768.46	27304.18	651.00	5100.00	7620.00	10083.00	871200.00

Fonte: Autoria própria.

Foi investigada a existência de duplicatas no banco de dados, para entender se haviam ou não linhas repetidas e apesar de um retorno negativo para esta questão, foram observados que alguns valores no campo *ID* se repetiam. Esse fato é consequência de uma mesma propriedade ter sido negociada em diferentes momentos como é o caso da propriedade de *ID* igual a 795000620 que foi negociada em três momentos conforme a Tabela 8.

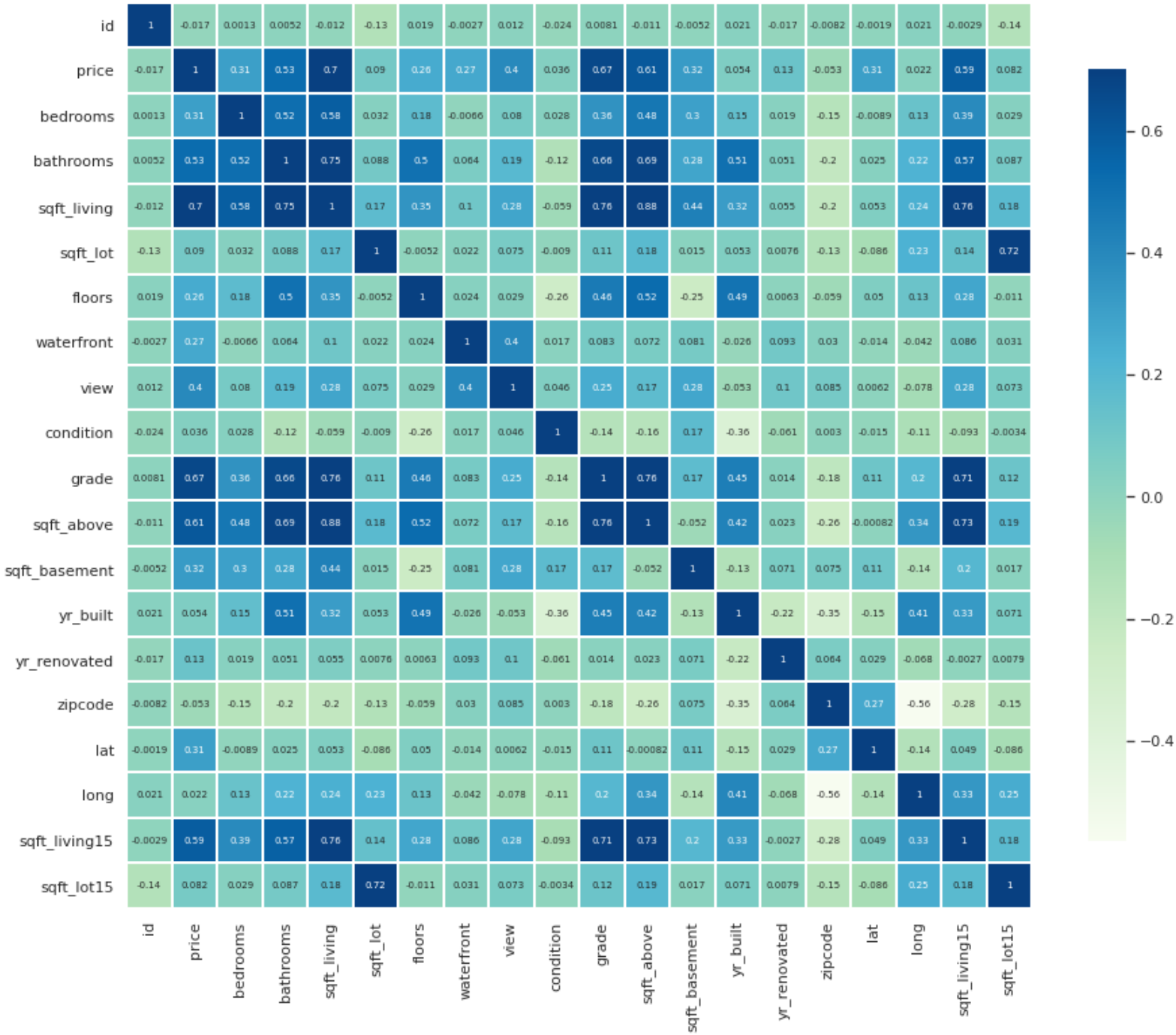
Tabela 8: Tabela trazendo exemplos de IDs repetidos.

id	price	bedrooms	bathrooms	date
795000620	115000.00	3	1.00	24/09/2014
795000620	124000.00	3	1.00	15/12/2014
795000620	157000.00	3	1.00	11/03/2015

Fonte: Autoria própria.

Para observar a correlação entre as variáveis do banco de dados foi gerada a Figura 1 que através de um mapa de calor possibilita uma comparação geral dos campos entre si.

Figura 1: Correlação de Pearson aplicada às variáveis pré tratamento de outliers.



Fonte: Autoria própria.

De todas as informações disponíveis na Figura 1, a correlação entre o preço e os demais recursos do imóvel é a mais importante. Para facilitar essa visualização foi construída a Tabela 9 onde os valores comparados apenas com a variável preço estão dispostos de forma decrescente permitindo observar diretamente as variáveis com correlação mais forte com o preço do imóvel.

Tabela 9: Lista de correlações em comparação com a variável “preço do imóvel” pré tratamento de outliers.

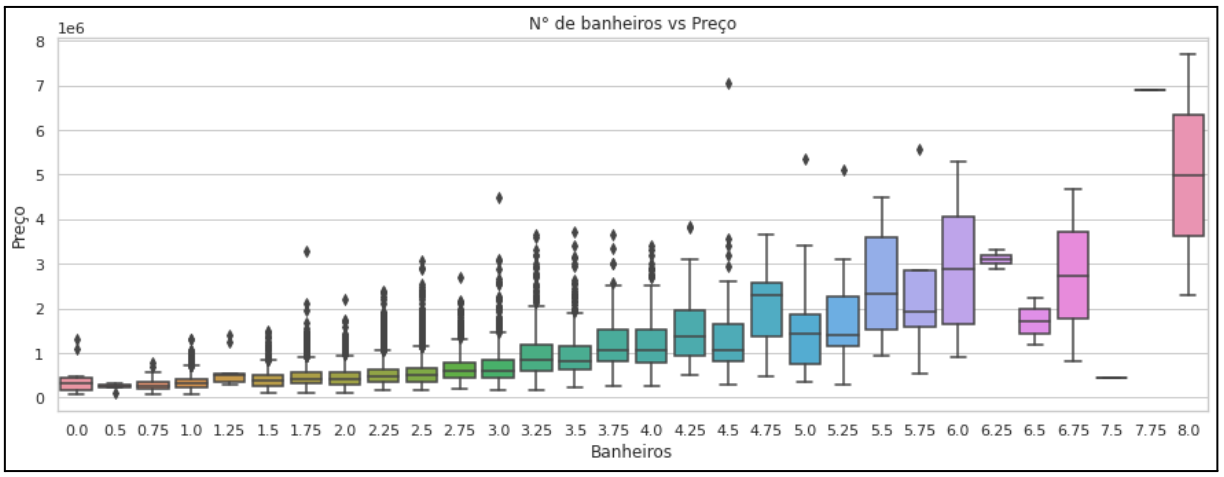
Campo	Correlação
price	1,00000000
sqft_living	0,70203500
grade	0,66743400
sqft_above	0,60556700
sqft_living15	0,58537900
bathrooms	0,52513800
view	0,39729300
sqft_basement	0,32381600
bedrooms	0,30835000
lat	0,30700300
waterfront	0,26636900
floors	0,25679400
yr_renovated	0,12643400
sqft_lot	0,08966100
sqft_lot15	0,08244700
yr_built	0,05401200
condition	0,03636200
long	0,02162600
zipcode	-0,05320300

Fonte: Autoria própria.

Concluiu-se a partir da correlação de Pearson que a metragem quadrada da área interna da construção, a nota em relação à construção/design e a metragem quadrada do interior da casa dos andares acima do térreo são os três itens que mais impactam no valor do imóvel. É importante observar que até este momento não foram feitos tratamentos referentes a *outliers*.

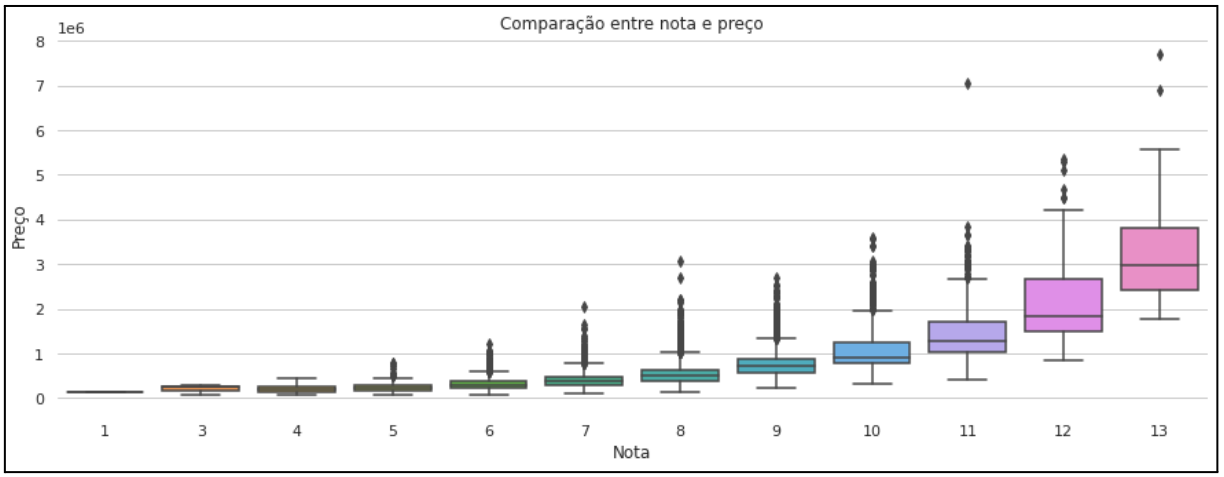
Nas Figuras 2, 3 e 4 foram construídos gráficos *boxplot* comparando o número de banheiros, o número de quartos e o valor da nota com o preço dos imóveis. Na Figura 2 temos a distribuição do número de imóveis em função do número de banheiros, sendo que os valores quebrados se referem a lavabos.

Figura 2: Boxplot do número de banheiros em relação ao preço, pré tratamento de outliers.



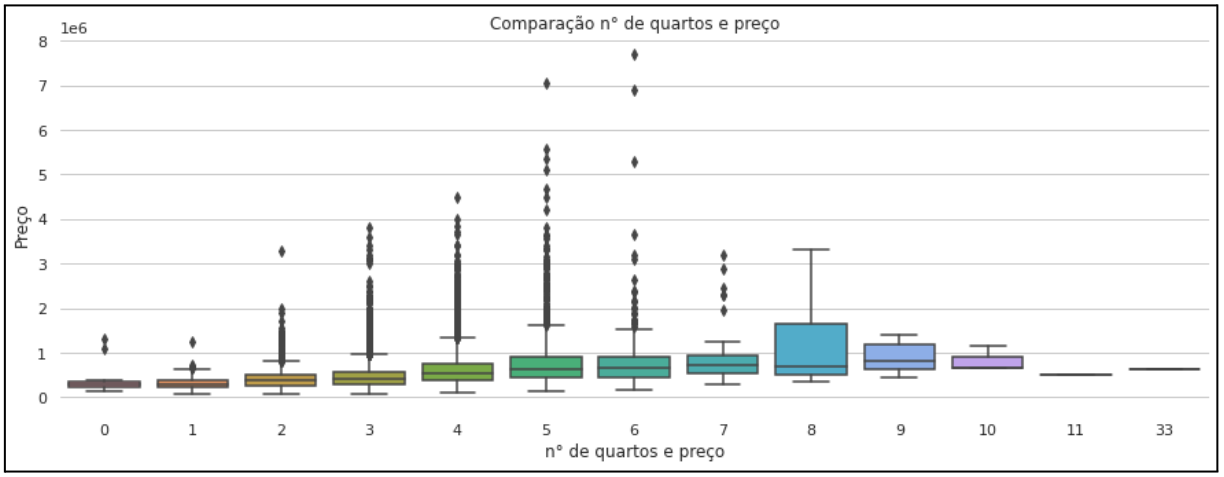
Fonte: Autoria própria.

Figura 3: Boxplot da nota em relação ao preço, pré tratamento de outliers.



Fonte: Autoria própria.

Figura 4: Boxplot do número de quartos em relação ao preço, pré tratamento de outliers.

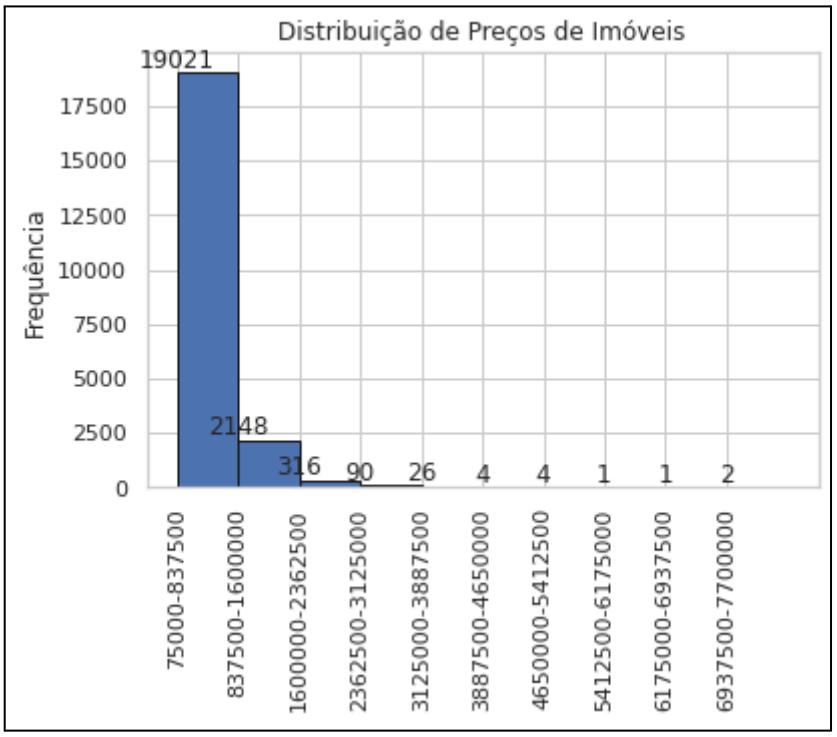


Fonte: Autoria própria.

Por meio do gráfico *boxplot*, é possível notar uma tendência de valorização imobiliária a partir do aumento do número de banheiros, número de quartos e valor da nota. Como observado nos gráficos boxplot existem outliers neste conjunto de dados. Essa dispersão pode influenciar a análise por apresentar muitos valores destoantes e bastante expressivos. Portanto foi feito um tratamento desses outliers onde procurou-se remover as maiores discrepâncias.

Para começar o tratamento dos dados foram escolhidas 6 variáveis para observar a distribuição dos valores e aplicar algum tratamento que no caso significou a remoção de alguns valores. Foi gerado o gráfico presente na Figura 5, onde após observar a distribuição do preço, foram removidos todos os imóveis que apresentavam um preço maior ou igual a R$1.600.000,00, o que significou uma redução de aproximadamente 2% da base de dados.

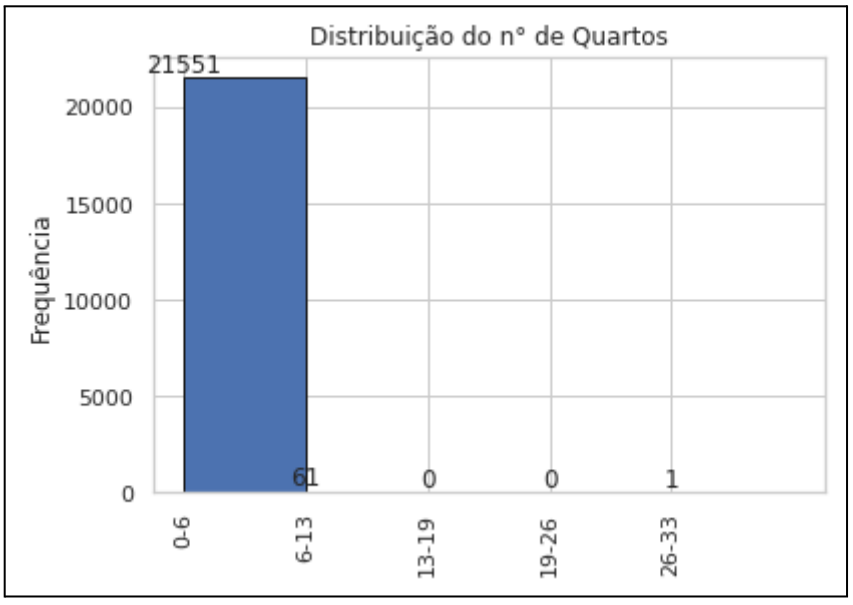
Figura 5: Distribuição dos valores dos imóveis.



Fonte: Autoria própria.

Em seguida, após observar o gráfico presente na Figura 6 foram removidos todos os imóveis que apresentavam número de quartos maior do que 6.

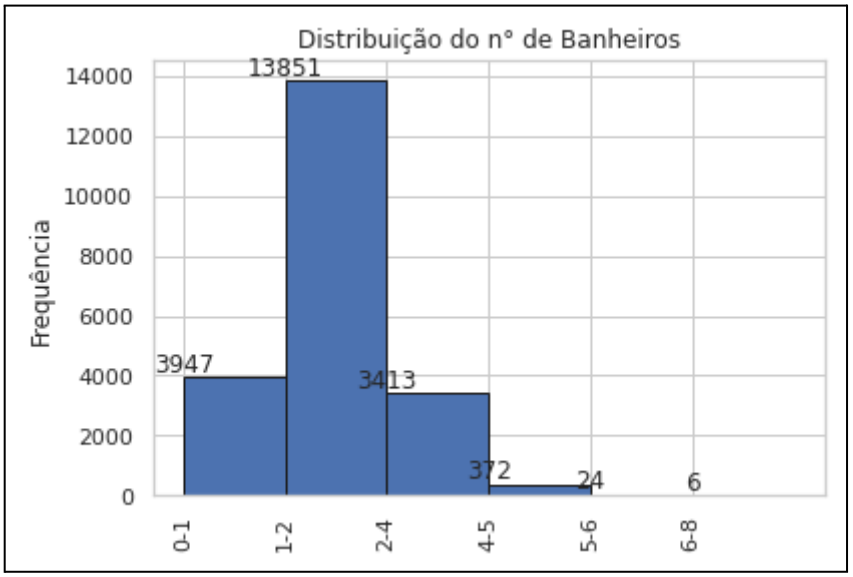
Figura 6: Distribuição do número de quartos dos imóveis.



Fonte: Autoria própria.

Após observar o gráfico presente na Figura 7 foram removidos todos os imóveis que apresentavam número de banheiros maior ou igual a 4.

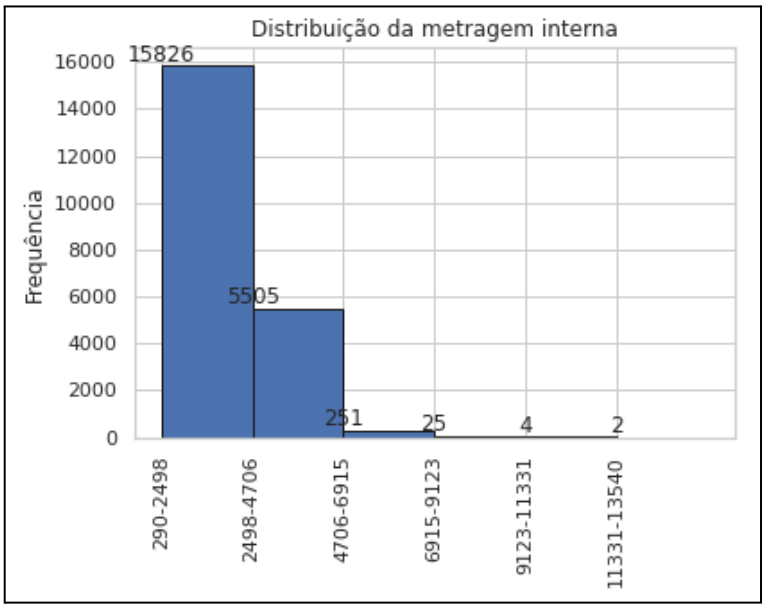
Figura 7: Distribuição do número de banheiros dos imóveis.



Fonte: Autoria própria.

Após observar o gráfico presente na Figura 8 foram removidos todos os imóveis que apresentavam metragem quadrada interna maior ou igual a 4706 m².

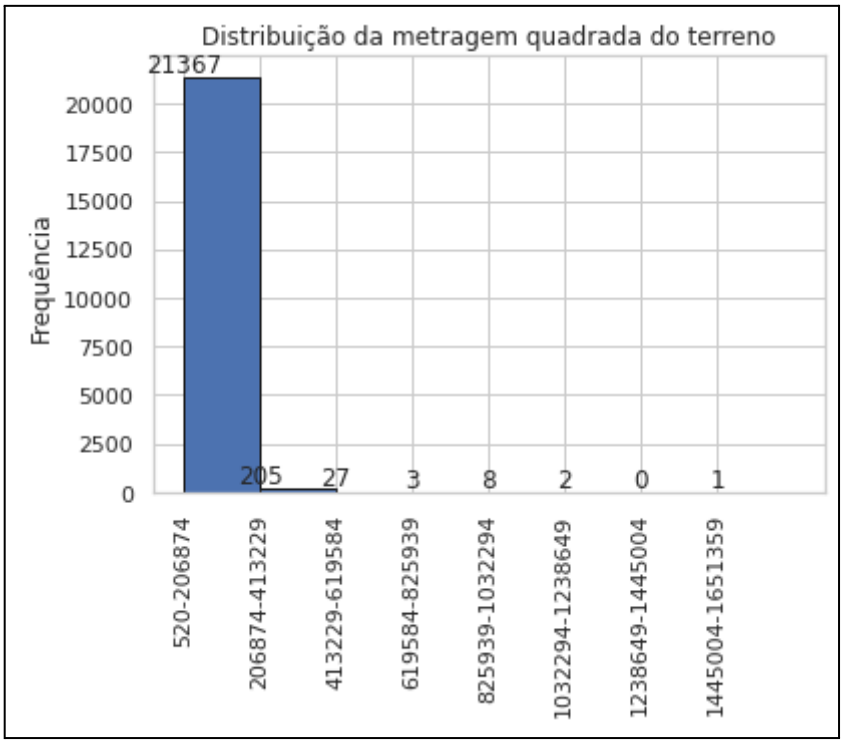
Figura 8: Distribuição da metragem quadrada interna dos imóveis.



Fonte: Autoria própria.

Após observar o gráfico presente na Figura 9 foram removidos todos os imóveis que apresentavam metragem quadrada do terreno maior ou igual a 206.874 m².

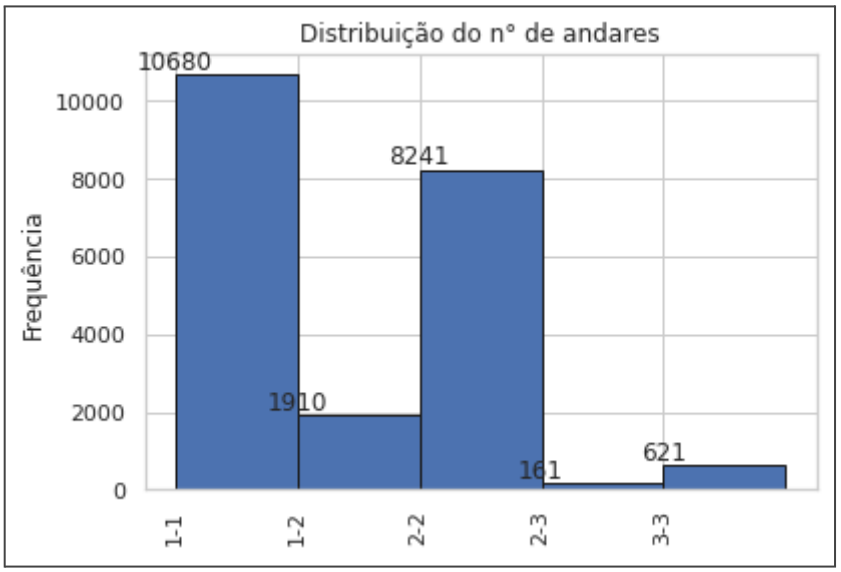
Figura 9: Distribuição da metragem quadrada do terreno dos imóveis.



Fonte: Autoria própria.

Por último, após observar o gráfico presente na Figura 10 foram removidos todos os imóveis que apresentavam número de andares maior ou igual a 3.

Figura 10: Distribuição do número de andares dos imóveis.



Fonte: Autoria própria.

Finalizada a remoção dos valores discrepantes restaram na base de dados 20.012 imóveis o que resultou em uma redução final de aproximadamente 7,4%. Ao gerar a Tabela 10 estamos apenas atualizando as informações da Tabela 7, uma vez que houve uma redução considerável no volume de dados e isso afetará as demais medidas estatísticas presentes na Tabela 10.

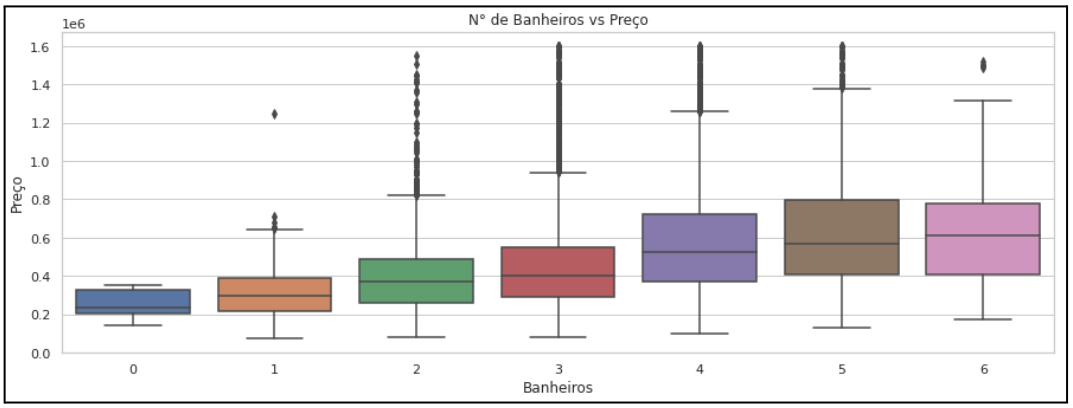
Tabela 10: Tabela descritiva do banco de dados com algumas medidas estatísticas pós tratamento de outliers.

Variáveis	média	desv. padrão	mínimo	25%	50%	75%	máximo
id	-	-	-	-	-	-	-
price	495583.66	248518.71	75000.00	315000.00	440000.00	618000.00	1600000.00
bedrooms	3.34	0.86	0.00	3.00	3.00	4.00	6.00
bathrooms	02.04	0.69	0.00	1.50	2.00	2.50	3.75
sqft_living	1994.92	771.85	290.00	1410.00	1890.00	2470.00	4700.00
sqft_lot	11640.21	17505.97	520.00	5120.00	7589.00	10350.00	206480.00
floors	1.43	0.48	1.00	1.00	1.00	2.00	2.50
waterfront	0.00	0.06	0.00	0.00	0.00	0.00	1.00
view	0.19	0.68	0.00	0.00	0.00	0.00	4.00
condition	3.42	0.66	1.00	3.00	3.00	4.00	5.00
grade	7.55	01.08	1.00	7.00	7.00	8.00	12.00
sqft_above	1715.82	726.14	290.00	1170.00	1530.00	2130.00	4700.00
sqft_basement	279.10	414.77	0.00	0.00	0.00	550.00	2330.00
yr_built	1969.51	29.04	1900.00	1950.00	1972.00	1994.00	2015.00
yr_renovated	80.37	392.36	0.00	0.00	0.00	0.00	2015.00
zipcode	98077.92	53.87	98001.00	98033.00	98065.00	98118.00	98199.00
lat	47.56	0.14	47.16	47.46	47.57	47.68	47.78
long	-122.21	0.14	-122.51	-122.33	-122.23	-122.13	-121.31
sqft_living15	1948.89	630.94	399.00	1480.00	1830.00	2320.00	5790.00
sqft_lot15	11006.74	18371.50	651.00	5175.00	7615.00	9910.00	434728.00
month	6.57	3.11	1.00	4.00	6.00	9.00	12.00
year	2014.32	0.47	2014.00	2014.00	2014.00	2015.00	2015.00

Fonte: Autoria própria.

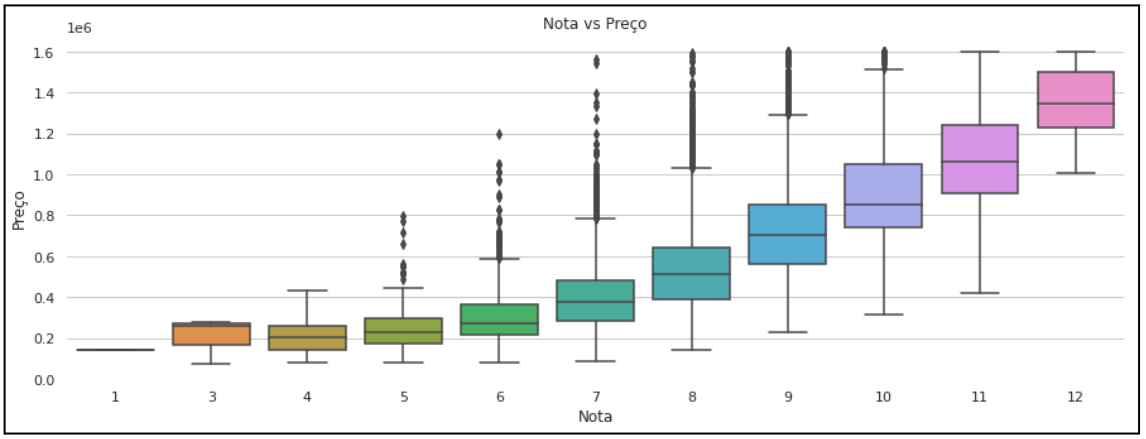
Através das Figuras 11, 12 e 13 pode-se observar as mesmas informações presentes nas Figuras 2, 3 e 4 porém agora atualizadas com a remoção dos valores mais discrepantes.

Figura 11: Boxplot do número de banheiros em relação ao preço, pós tratamento de outliers.



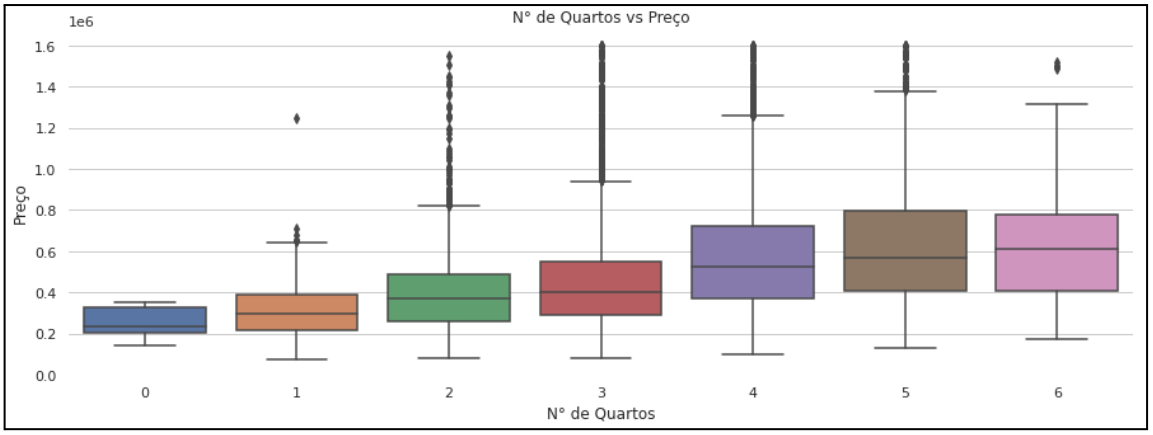
Fonte: Autoria própria.

Figura 12: Boxplot da nota em relação ao preço, pós tratamento de outliers.



Fonte: Autoria própria.

Figura 13: Boxplot do número de quartos em relação ao preço, após tratamento de outliers.



Fonte: Autoria própria.

Através dos gráficos *boxplot*, é possível notar que ainda assim há uma tendência de valorização imobiliária a partir do aumento do número de banheiros, número de quartos e valor da nota. Para responder então à primeira pergunta de negócios foi gerada a Tabela 11 que atualiza os dados da Tabela 9 após a remoção dos outliers.

Tabela 11: Lista de correlações em comparação com a variável “preço do imóvel” pós tratamento de outliers.

Campo	Correlação
price	1.00
grade	0.65
sqft_living	0.64
sqft_living15	0.58
sqft_above	0.53
bathrooms	0.46
lat	0.41
view	0.31
bedrooms	0.30
floors	0.29
sqft_basement	0.27
yr_renovated	0.11
waterfront	0.10
sqft_lot	0.09
sqft_lot15	0.06
condition	0.05
yr_built	0.03
long	0.03
id	0.00
year	0.00

Fonte: Autoria própria.

Concluiu-se a partir da correlação de Pearson que a nota em relação à construção/design, metragem quadrada da área interna da construção, a metragem quadrada média do interior da construção dos 15 vizinhos mais próximos, a metragem quadrada do interior da casa dos andares acima do térreo e o número de banheiros são os recursos que mais impactam no valor do imóvel em ordem decrescente.

A segunda pergunta a ser respondida é: a House Rocket poderia fazer uma reforma para aumentar o preço da venda? Quais seriam as sugestões de mudanças?

Observando a Tabela 11, os primeiros itens são os de maior correlação com a variável preço e estes serão consequentemente aqueles que têm maior potencial em aumentar ou diminuir o valor de um imóvel se alterados. A começar pela variável *bathrooms* que representa o número de banheiros, além da correlação de Pearson, foi possível observar através da Figura 11 que há de fato uma tendência valorização do imóvel à medida que o número de banheiros aumenta. Dessa forma, aumentar o número de banheiros na propriedade é uma das opções de reforma.

As variáveis *sqft_above*, *sqft_living15* e *sqft_living* são as próximas variáveis que apresentam maior correlação a partir da Tabela 11, porém é importante observar que a metragem quadrada média do interior da construção dos 15 vizinhos mais próximos não é uma variável que pode ser alterada pelo proprietário no caso de uma reforma, portanto a variável *sqft_living15* é excluída das possibilidades. No caso de *sqft_above* e *sqft_living* a se possível podem ser feitas reformas para aumentar a metragem quadrada tanto da construção interna como um todo quanto da metragem quadrada dos andares acima do térreo caso existam.

Por último, a nota do imóvel tem a maior correlação com o preço dos imóveis segundo a Tabela 11 e portanto é válido buscar aumentar o valor da nota para que consequentemente o imóvel seja valorizado. Porém, não é possível alterar diretamente a nota do imóvel, uma vez que esta é resultado de uma avaliação feita a partir de todos os recursos presentes e, portanto, num primeiro momento não indica uma ação direta a ser realizada. Dessa forma, foi analisada a correlação da nota do imóvel (*grade*) com as demais variáveis para compreender quais recursos poderiam contribuir mais com o aumento da nota do imóvel.

Tabela 12: Lista de correlações em comparação com a variável “nota do imóvel”, pós tratamento dos outliers.

Campo	Correlação
grade	1.00
sqft_living	0.73
sqft_above	0.72
sqft_living15	0.70
price	0.65
bathrooms	0.63
floors	0.48
yr_built	0.47
bedrooms	0.35
long	0.22
view	0.17
sqft_lot	0.13
sqft_lot15	0.10
lat	0.10
sqft_basement	0.10
id	0.03
month	0.01
waterfront	0.00
yr_renovated	-0.00
year	-0.04
condition	-0.15
zipcode	-0.19

Fonte: Autoria própria.

As variáveis *sqft_living*, *sqft_above* e *bathrooms* são os três recursos passíveis de alteração mediante a uma reforma que apresentam maior correlação com a variável *grade*, indicando portanto que as sugestões realizadas anteriormente já terão impacto também na nota do imóvel. Há ainda a variável *bedrooms* que apresenta uma correlação moderada de acordo com a Tabela 4 e que apresenta também uma tendência de valorização da propriedade à medida que o número de quartos aumenta conforme observado na Figura 13.

5 CONSIDERAÇÕES FINAIS

Neste estudo foram utilizadas diferentes ferramentas estatísticas de forma bem-sucedida para explorar, compreender e extrair insights importantes a fim de responder às perguntas de negócio propostas.

Ao investigar quais características de um imóvel mais impactam no seu preço, foi possível identificar, por meio da correlação de Pearson, que a área de construção interna, a área de construção acima do térreo, a nota do imóvel e o número de banheiros são as características que mais afetam o preço. Além disso, ao combinar a correlação com a visualização de dados por meio de gráficos de dispersão, *boxplot* e linhas de tendência, observou-se que reformar um imóvel para melhorar a área de construção interna, área de construção acima do térreo ou o número de banheiros pode ser uma estratégia eficaz para valorizar o imóvel.

Conclui-se então que mesmo utilizando poucas ferramentas estatísticas é possível realizar uma análise exploratória dos dados a fim de compreender e analisar as informações disponíveis a fim de extrair insights.

REFERÊNCIAS

KAGGLE. **Kaggle**. Disponível em: http://www.kaggle.com. Acesso em: 13 ago. 2023.

SANTOS JUNIOR, Anísio Pereira dos. **Análise das características de jogabilidade no PUBG usando árvore de decisão**. 2019. 26 f. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal de Uberlândia, Uberlândia, 2020. Disponível em: https://repositorio.ufu.br/bitstream/123456789/28354/7/An%c3%a1liseCaracter%c3%adsticasJogabilidade.pdf. Acesso em: 27 fev. 2023.

MARRA, Vinicius Nogueira. **Previsão de dificuldades financeiras em empresas latino-americanas via aprendizagem de máquina**. 2019. 81 f. Dissertação (Mestrado em Administração) - Universidade Federal de Uberlândia, Uberlândia, 2019. Disponível em: https://repositorio.ufu.br/bitstream/123456789/24750/1/Previs%c3%a3oDificuldadesFinanceiras.pdf. Acesso em: 27 fev. 2023.

VAZ, Hélida Pereira. **Regressão Linear na avaliação do desempenho escolar de alunos do 3º ano do Ensino Fundamental**. 2016. 18 f. Trabalho de Conclusão de Curso (Graduação em Matemática) - Universidade Federal de Uberlândia, Uberlândia, 2020. Disponível em: https://repositorio.ufu.br/bitstream/123456789/29070/1/Regress%c3%a3oLinearAvalia%c3%a7%c3%a3o%20.pdf. Acesso em: 27 fev. 2023.

FREITAS, Pablo Henrique de. **Regressão Logística na modelagem da probabilidade de vitória em jogos de futebol americano**. 2019. 54 f. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal de Uberlândia, Uberlândia, 2019. Disponível em: https://repositorio.ufu.br/bitstream/123456789/26405/4/Regress%c3%a3oLog%c3%adsticaModelagem.pdf. Acesso em: 01 jan. 2023.

BUSSAB, W. O.; MORETTIN, P. A.. **Estatística Básica**. 6 ed. São Paulo: Saraiva, 2010.

LUCASCAPOVILLA. **House Rocket**: Notebook. Disponível em: https://www.kaggle.com/code/lucascapovilla/house-rocket. Acesso em: 13 ago. 2023.

LEVIN, Jack; FOX, James A.; FORDE, David R.. **Estatística para Ciências Humanas**. Tradução Jorge Ritter. 11ª ed. São Paulo: Pearson Education do Brasil, 2012.

PYTHON. **Python Software Foundation**. Disponível em: https://www.python.org/. Acesso em: 13 ago. 2023.

HOTZ, Nick. **What is CRISP DM?**. 2023. Disponível em: https://www.datascience-pm.com/crisp-dm-2/. Acesso em: 13 dez. 2022.

PANDAS, Development Team. **Biblioteca Pandas**. 2023. Disponível em: https://pandas.pydata.org/. Acesso em: 15 out. 2023.

MATPLOTLIB, Development Team. **Biblioteca Matplotlib**. 2023. Disponível em: https://matplotlib.org/index.html. Acesso em: 15 out. 2023.

SEABRON, Development Team. **Biblioteca Seaborn**. 2023. Disponível em: https://seaborn.pydata.org/. Acesso em: 15 out. 2023.