Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO EM

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

PROJETO

UTILIZAÇÃO DE MODELOS DE APRENDIZAGEM AUTOMÁTICA PARA PREVISÃO DE PREÇO DE AIRBNB EM LISBOA

DINIS DUARTE SALGUEIRINHO

ORIENTAÇÃO:

PROF. DOUTOR CARLOS J. COSTA

OUTUBRO-2023

Agradecimentos

À minha família, pelo apoio que me deram em todas as fases da minha vida, que se revelou crucial para a elaboração deste projeto final de Mestrado. Agradeço a compreensão e motivação com que pude contar ao longo deste trabalho.

Ao meu orientador, Prof. Doutor Carlos Costa, por todo o apoio e paciência que teve comigo. Agradeço-lhe pela sua boa disposição, compreensão, conselhos e total disponibilidade que revelou em todas as fases do trabalho e, acima de tudo, pela confiança que depositou em mim.

Aos meus amigos, pela força que sempre me transmitiram para elaborar o projeto e pela total compreensão da minha ausência, em muitos dos eventos de confraternização que tiveram lugar durante este período.

Por último, mas (certamente) não menos importante, gostaria de estender o meu mais profundo agradecimento a mim mesmo pela dedicação incansável, pela perseverança e pela determinação ao longo deste desafiante percurso académico. Este trabalho representa não apenas a minha investigação, mas também o meu compromisso com o crescimento pessoal e o desejo constante de aprender e evoluir.

Resumo

A economia de partilha surgiu há cerca de vinte anos. Era difícil prever a sua importância para muitos setores, mas, atualmente, está intimamente incorporada no nosso quotidiano. O Airbnb é um excelente exemplo deste fenómeno, uma vez que introduziu um novo modelo de negócio para o sector da hotelaria. Deste modo, prever com precisão uma variável contínua, como o preço, em anúncios do Airbnb tem sido um tema de elevada importância.

Este documento analisa uma amostra de 19651 anúncios em Lisboa do InsideAirbnb.com, com o objetivo de criar um modelo de previsão de preços com recurso a técnicas de aprendizagem automática.

Inicialmente, é realizada uma limpeza e pré-processamento dos dados, seguida de análises descritivas, prescritivas e exploratórias para compreender a natureza dos dados e identificar atributos importantes para a previsão de preços. Mesmo após a limpeza, valores atípicos são detetados e removidos do conjunto de dados. Por fim, aplicámos modelos que vão desde a regressão linear a modelos de Machine Learning. Desta forma, não só selecionámos o XGBoost como o melhor modelo para a previsão de preços com R^2 de 0.6225, mas também identificamos as características que têm significância estatístico para a variável dependente preço.

Palavras-chave: Aprendizagem automática; Airbnb; Previsão de preços; Regressão Linear; XGBRegressor

Abstract

The sharing economy emerged just twenty years ago. It was difficult to foresee its importance for many sectors, but today it is intimately incorporated into our daily lives. Airbnb is an excellent example of this phenomenon, as it has introduced a new business model for the hospitality sector. Therefore, accurately predicting a continuous variable such as price in Airbnb listings has been a topic of significant importance.

This paper analyses a sample of 19651 listings in Lisbon from InsideAirbnb.com, with the aim of creating a price prediction model using machine learning techniques.

Initially, the data is cleaned and pre-processed, followed by descriptive, prescriptive, and exploratory analysis to understand the nature of the data and identify important attributes for price prediction. Even after cleaning, outliers are detected and removed from the data set. Finally, we applied models ranging from linear regression to machine learning models. In this way, we not only selected XGBoost as the best model for price prediction with an R2 of 0.6225, but also identified the characteristics that have statistical significance for the dependent variable price.

Keywords: Machine Learning; Airbnb; Price Prediction; Linear Regression; XGBRegressor

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

1.1. Enquadramento

A economia de partilha, também conhecida como economia colaborativa ou economia partilhada, é um conceito que se refere a um modelo económico em que a propriedade e o acesso a bens e serviços são partilhados entre indivíduos ou grupos de pessoas, frequentemente facilitados por plataformas digitais. Esse modelo contrasta com a economia tradicional, em que a posse exclusiva de bens e a prestação de serviços geralmente são centralizadas em empresas.

A economia de partilha é caracterizada por uma série de princípios fundamentais que moldam sua natureza e funcionamento. Esses princípios incluem a prevalência do acesso sobre a propriedade, onde as pessoas optam por partilhar o acesso a recursos, em vez de adquiri-los (Polisetty & Kurian, 2021). Evidencia-se um papel essencial das plataformas digitais, que atuam como intermediárias, que interligam fornecedores de serviços e proprietários de ativos a consumidores interessados, tornando o partilha mais acessível e eficiente. Por outro lado, a promoção do uso eficiente de recursos, uma vez que os bens e serviços são partilhados, reduz o desperdício e maximiza a utilização dos mesmos. A ênfase na confiança, através de avaliações e comentários, desempenha um papel vital na construção da reputação e garante a integridade deste modelo económico. A convergência com os princípios da economia circular promove a reutilização e o aumento do ciclo de vida dos bens, fundamental para a sustentabilidade e a redução do impacto ambiental. Esses princípios da economia de partilha não apenas influenciam a maneira como as transações económicas são conduzidas, mas também têm um impacto mais amplo na cultura de consumo e na gestão de recursos. A economia de partilha tem crescido rapidamente em todo o mundo, desafiado os modelos de negócios tradicionais em muitos setores (Polisetty & Kurian, 2021).

O Airbnb está intrinsecamente relacionado com a economia de partilha, uma vez que representa um dos exemplos mais proeminentes e bem-sucedidos desta abordagem económica. O Airbnb é uma plataforma online que permite que as pessoas encontrem alojamento, como casas, apartamentos, quartos individuais e até mesmo espaços únicos, como castelos e casas na árvore, para estadias curtas ou prolongadas. Fundada em agosto de 2008 por Brian Chesky, Joe Gebbia e Nathan Blecharczyk, a empresa revolucionou o modelo de negócio em torno do alojamento temporário (Aydin, 2019).

O Airbnb facilita a ligação entre anfitriões, que são os proprietários ou arrendatários dos imóveis, e potenciais hóspedes. Os anfitriões podem publicar as suas propriedades na plataforma, descrevendo detalhadamente as suas características, incluindo fotos, preços e regras da casa. Por outro lado, os potenciais hóspedes podem pesquisar as várias ofertas, ler avaliações e comentários de outros hóspedes, fazer reservas e pagar diretamente através do *website* ou da aplicação móvel (Zekanovic-Korona, & Grzunov, 2014). A crescente popularidade do Airbnb desde a sua criação tornou a empresa uma das líderes no setor de alojamento. Esse sucesso pode ser atribuído a diversos fatores-chave que caracterizam a plataforma (Zekanovic-Korona, & Grzunov, 2014).Uma ampla variedade de opções de alojamento é disponibilizada pelo Airbnb, abrangendo desde quartos privados em casas partilhadas até mansões de luxo. Esta diversidade de oferta dá resposta a diferentes orçamentos e preferências. Os preços oferecidos pelo Airbnb revelam-se frequentemente mais acessíveis em comparação com os de hotéis tradicionais, especialmente para viagens de grupo ou estadias prolongadas. (Thaichon, et al.2020).

A experiência proporcionada pelo Airbnb está mais enraizada na cultura local, uma vez que, permite aos hóspedes ficarem alojados em bairros residenciais promovendo a interação com a comunidade. As avaliações e feedback dos hóspedes são uma parte fundamental da plataforma, pois permitem que, quem esteja interessado possa tomar decisões informadas. Essas avaliações ajudam a garantir a qualidade da experiência (Thaichon, et al.2020). A flexibilidade oferecida pelo Airbnb é evidente na possibilidade de os hóspedes escolherem datas de check-in e check-out, proporcionando maior conveniência. Além dos benefícios para quem procura alojamento, o Airbnb proporciona uma oportunidade de negócio para os anfitriões, que podem rentabilizar as suas propriedades ou quartos vagos (Thaichon, et al.2020).

Em resumo, o êxito do Airbnb deve-se a uma combinação destes fatores, criando uma plataforma que oferece diversidade, preços competitivos, experiência autêntica, confiança e flexibilidade (Thaichon, et al.2020). O Airbnb continua a ser uma influência relevante no setor de alojamento e um exemplo proeminente da economia de partilha em prática. No entanto, o Airbnb também enfrentou desafios e controvérsias ao longo dos anos. Questões relacionadas com regulamentações, o impacto no mercado de arrendamento a longo prazo e as preocupações sobre a segurança dos hóspedes e a integridade das propriedades estão entre os tópicos mais debatidos (Cassell & Deutsch, 2020).

A partilha de casa, em particular, tem sido objeto de intensas críticas. Nomeadamente, os críticos argumentam que as plataformas de partilha de casas, como a Airbnb, aumentam o custo de vida dos inquilinos locais, beneficiando sobretudo os proprietários e os turistas não residentes. É fácil perceber o argumento económico. A redução das barreiras no mercado de

alojamento de curta duração por meio das plataformas digitais, tem como consequência a migração de alguns senhorios do mercado de arrendamentos a longo prazo, no qual é mais provável que estejam envolvidos residentes locais, para o mercado de alojamento de curta duração, no qual a participação de não residentes é mais comum (Cassell & Deutsch, 2020). Uma vez que a oferta total de habitação é fixa ou inelástica a curto prazo, este facto faz aumentar o custo de arrendamento no mercado de longo prazo. A preocupação com o impacto da partilha de casa na acessibilidade da habitação tem merecido grande atenção por parte dos decisores políticos e motivou já muitas cidades a imporem regulamentação mais rigorosa (Cassell & Deutsch, 2020).

Em resumo, o Airbnb é uma plataforma disruptiva que mudou a maneira como as pessoas viajam e encontram alojamento temporário. A sua história de sucesso é resultado da combinação de inovação tecnológica com a diversidade de opções de alojamento e a capacidade de proporcionar experiências únicas. No entanto, também levanta questões complexas sobre regulamentação, concorrência e responsabilidade que continuam a ser debatidas em âmbito global.

1.2. Objetivos

O mercado de alojamento de curta duração, representado em grande parte por plataformas como o Airbnb, tem desempenhado um papel significativo na indústria do alojamento. A dinâmica deste mercado é influenciada por uma série de fatores, tornando a previsão de preços uma tarefa desafiadora, mas crucial para os anfitriões e para os hóspedes.

Este projeto de investigação concentra-se na previsão de preços de anúncios do Airbnb em Lisboa, explorando os fatores que afetam as decisões de preços dos anfitriões e as expectativas dos hóspedes. Compreender os determinantes dos preços e desenvolver modelos de previsão precisos é de grande importância para todos os envolvidos neste ecossistema.

Como tal definimos como objetivos deste projeto os seguintes pontos:

Identificar os Principais Determinantes de Preços: Investigar e analisar os fatores mais influentes que moldam os preços dos anúncios no Airbnb em Lisboa. Isso inclui a análise de variáveis como a localização, o tipo de propriedade, as comodidades oferecidas, a disponibilidade e outros fatores que possam desempenhar um papel fundamental na formação dos preços.

Avaliar o Impacto das Avaliações dos Hóspedes: Examinar como as avaliações dos hóspedes, incluindo as avaliações especificas e os comentários, impactam os preços em Lisboa.

O objetivo é entender como a satisfação do hóspede influencia as estratégias de preços adotadas pelos anfitriões.

Comparar Diferentes Modelos de Previsão: Comparar a eficácia de diversos modelos de previsão, tais como regressão linear, árvores de decisão, redes neuronais (NN) e outros, a fim de determinar qual o modelo mais adequado para a previsão de preços no contexto do Airbnb em Lisboa.

Investigar Tendências de Preços: Explorar as tendências de preços ao longo do tempo em Lisboa.

Estes objetivos orientam a nossa investigação, a recolha e a análise de dados, bem como a construção de modelos de previsão, contribuindo para um entendimento mais aprofundado dos fatores que influenciam os preços no mercado de alojamento de curta duração em Lisboa.

1.3. Abordagem metodológica

A ciência dos dados tem vindo a assumir uma importância crescente na nossa sociedade, impulsionando a tomada de decisões informadas, a resolução de problemas complexos e a exploração de oportunidades de inovação (Aparicio et al., 2019, Mergulhão, et al. 2022, Costa & Aparicio, 2023). O acesso a grandes volumes de dados e os avanços nas tecnologias de informação abriram novas perspetivas para a aquisição de conhecimento e a compreensão de diversos fenómenos (Arriaga & Costa, 2022).

Neste contexto, a ciência dos dados emergiu como uma disciplina fundamental, uma vez que permite extrair conhecimento com base num conjunto de dados complexos, revelando padrões, tendências e *insights* que por vezes não são facilmente percetíveis através da observação direta. No entanto, a ciência dos dados não é uma tarefa simples. Requer uma abordagem sistemática e bem estruturada para assegurar a validade e a utilidade dos resultados obtidos (Costa & Aparicio, 2020, 2021, Tavares, et al, 2022).

É neste ponto que entra em jogo a abordagem metodológica CRISP-DM (*Cross-Industry Standard Process for Data Mining*). A metodologia CRISP-DM é um tipo de metodologia que oferece orientações sólidas para uma eficaz e eficiente gestão de projetos de análise de dados. A sua estrutura abrange todas as fases essenciais de um projeto de análise de dados, desde a compreensão do problema até à implementação das soluções (Costa & Aparicio, 2020, 2021).

Este modelo é composto por seis fases interligadas, cada uma desempenhando um papel distintivo e crucial na construção do nosso projeto de análise de dados.

A primeira fase, "Compreensão do Negócio" representa o ponto de partida. Aqui, os objetivos de negócio são claramente definidos e os critérios de sucesso são estabelecidos. É essencial compreender o contexto do problema e os requisitos do projeto antes de prosseguir para as etapas técnicas.

A segunda fase, "Compreensão dos Dados" concentra-se na aquisição de dados relevantes e na exploração dos mesmos. A qualidade e a adequação dos dados são analisadas, para que seja possível compreender de modo mais aprofundado o conteúdo. Esta etapa é vital para que se faça uma seleção apropriada das variáveis e que se garanta que os dados são representativos do problema.

A terceira fase, "Preparação dos Dados" é crucial para a eficácia das fases subsequentes. Nesta etapa, os dados são limpos, transformados e selecionados conforme necessário. A qualidade dos modelos depende da qualidade dos dados de entrada, tornando esta fase de preparação fundamental.

Na quarta fase, "Modelação" entramos no cerne do processo de análise de dados. Modelos estatísticos ou de aprendizagem automática são desenvolvidos com base nos dados preparados. Diversas técnicas, como regressão, árvores de decisão, redes neuronais (NN) e outros algoritmos, podem ser aplicados. Os modelos são ajustados e avaliados para garantir que dão resposta aos objetivos do projeto.

A quinta fase, "Avaliação" envolve a análise crítica dos modelos gerados na fase de modelagem. Os modelos são avaliados em relação aos critérios de sucesso previamente definidos na fase de Compreensão do Negócio. Métricas de desempenho, como precisão, sensibilidade e especificidade, são frequentemente usadas para determinar a eficácia dos modelos.

Por fim, a sexta fase, "Implementação" diz respeito à integração dos resultados de análise de dados nas operações do negócio. Os modelos desenvolvidos são usados para tomar decisões, automatizar processos ou fornecer *insights* para o suporte à tomada de decisões. Esta fase procura garantir que o benefício do processo de análise de dados se traduz em ações reais e com impacto positivo no negócio (Aparicio, et al, 2023).

É importante realçar que o CRISP-DM é um processo iterativo, como é possível visualizar na figura 1. Isto significa que, ao longo do desenvolvimento do projeto, é possível regressar a fases anteriores para refinar os modelos, abordar novos desafios ou considerar novas perguntas. Essa flexibilidade é uma das razões para a ampla adoção do CRISP-DM na área de análise de dados (Aparicio, et al, 2023).

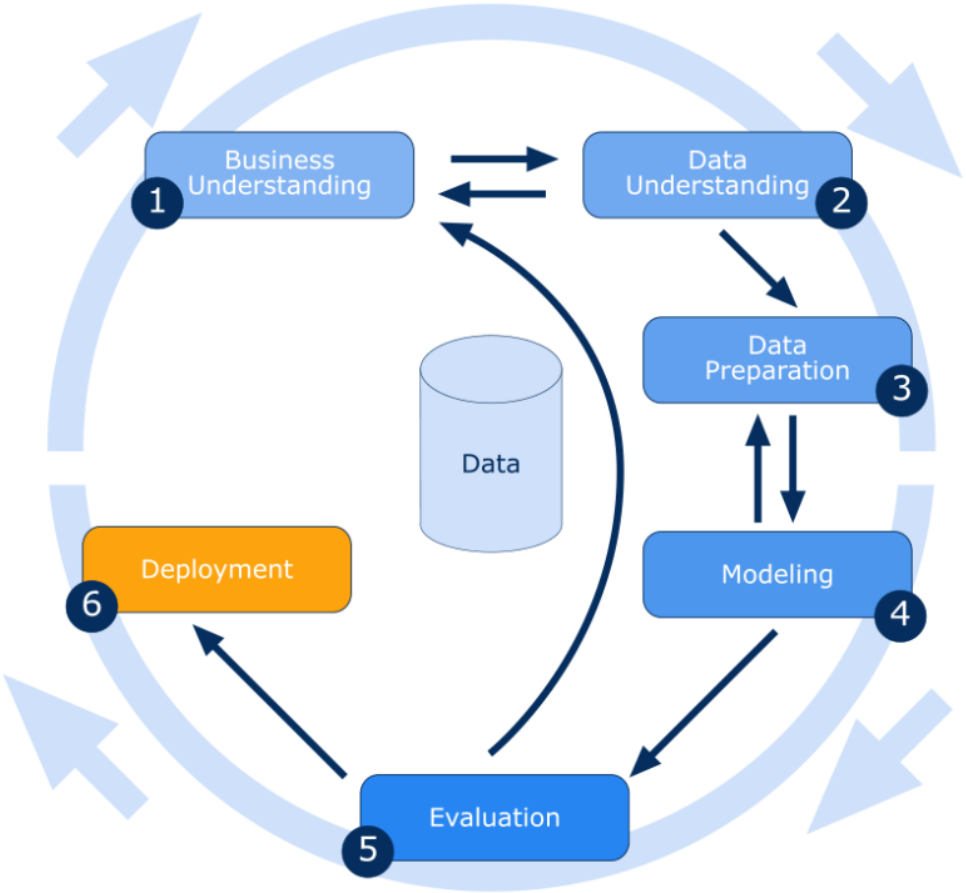


Figura 1- Diagrama CRISP-DM

Fonte: (Chapman et al., 2000)

1.4. Estrutura do relatório

A introdução é o ponto de partida do relatório, fornecendo uma visão geral do projeto. Inclui o enquadramento, que contextualiza o estudo e justifica a importância da previsão de preços no Airbnb; os objetivos do projeto, com uma linha clara sobre os objetivos deste projeto; a abordagem metodológica, descrevendo a metodologia utilizada no projeto, que será complementada no capítulo 3; e a estrutura do relatório, descreve uma visão geral das seções subsequentes.

Na seção, revisão de literatura são revistas as principais teorias, conceitos e trabalhos relacionados com previsão de preços no Airbnb.

O capítulo 3 contém a metodologia. Esta é uma seção crucial que descreve em detalhes como o projeto foi conduzido. Inclui a recolha e tratamento de dados, descrevendo as fontes de dados, processos de recolha e todas as técnicas de tratamento de dados; a análise exploratória de dados, com detalhes sobre como os dados foram explorados, visualizados e preparados para

a modelação; e a modelação, onde explicamos as técnicas de modelação utilizadas para prever os preços no Airbnb.

De seguida temos os resultados e a discussão onde são apresentados os resultados obtidos a partir da análise dos dados e da aplicação dos modelos de previsão. A discussão envolve a interpretação dos resultados, a contextualização em relação à literatura analisada e a análise das implicações dos resultados.

A conclusão resume as descobertas-chave do estudo, destaca quaisquer limitações encontradas e identifica possíveis áreas para trabalhos futuros.

Esta estrutura de relatório proporciona uma organização clara e informativa do trabalho desenvolvido sobre a previsão de preços no Airbnb.

2 Revisão de Literatura

Com o crescimento da economia de partilha, a plataforma Airbnb, que obtem receita através da cobrança a hóspedes e anfitriões por alojamentos de curta duração, está a tornar-se cada vez mais popular. A rápida expansão do Airbnb tem gerado um interesse significativo no meio académico. Por exemplo, Xie & Kwok (2017) investigaram o impacto do Airbnb nas receitas dos hotéis. O mercado imobiliário tem sido objeto de estudo tendo em conta diversas dimensões (Sadamani & Costa, 2021). Porém o impacto do Airbnb tem levado os investigadores a analisar a sua influência no mercado imobiliário (Bao & Shah, 2020, Garcia-López et. al., 2020). Autores ainda analisam os impactos em termos de gentrificação (Yrigoy, 2016). O preço, que influencia tanto os inquilinos como os proprietários das casas, não deve ser negligenciado (Zhang, et al. 2017). Dado que o Airbnb é uma plataforma onde as reservas são feitas com antecedência, ao contrário do pagamento à chegada como nos hotéis, é crucial que quem empreende no Airbnb, adote uma estratégia de preços que permita prever com precisão, de modo, a oferecer preços competitivos para atrair mais inquilinos (Zhang, et al. 2017, Kwok, & Xie, 2019).

Tradicionalmente, o Airbnb sugere aos proprietários que fixem inicialmente o preço com base nos valores das casas circundantes ou de casas semelhantes, aumentando os preços, ou até mesmo duplicando-o durante os períodos de férias, e baixando os preços quando se aproxima a data do check-in, caso a casa não tem sido reservada (Zhang, et al. 2017). No entanto, estas estratégias não consideram a enorme flexibilidade do mercado imobiliário, o que pode dificultar o arrendamento. Na verdade, são vários os fatores que podem influenciar os preços das casas na Airbnb. (Yrigoy, 2016, Zhang, et al. 2017, d'Orei Pape et al., 2022).

Assim, para garantir a estabilidade e a adequação dos preços da habitação, é fundamental adotar uma estratégia de preços que vá ao encontro da realidade. Muitos investigadores já se debruçaram neste campo, e vários modelos econométricos e de aprendizagem automática têm sido aplicados com sucesso (Zhang, et al. 2017, Gibbs, et. al., 2018).

Luo et al. (2019) desenvolveram um modelo de previsão de preços do Airbnb, utilizando *Random Forest, XGBoost* e Redes Neuronais (NN). Este processo partiu de uma seleção cuidadosa das características, excluindo algumas, como índice de *host* (host_id) e nome do cliente, para reduzir o ruído, enquanto mantiveram características como o código do país e o número de quartos. Além disso, consideraram variáveis de tipo texto, como a descrição da casa e as avaliações. Após uma extensa e meticulosa análise dos dados observou-se que XGBoost e

a Rede Neuronal (NN) apresentaram um desempenho superior em comparação com outros modelos.

Kalehbasti et al., (2021) apresentaram um modelo de previsão de preços sólido, baseado em técnicas de *Machine Learning*, *Deep learning* e processamento de linguagem natural, destinado a auxiliar ambas as partes, tanto o anfitrião como o consumidor, na tomada de decisões. O modelo de previsão proposto leva em consideração vários fatores de previsão, tais como as características da propriedade, do proprietário e as avaliações. Esses fatores são utilizados em algoritmos de aprendizagem automática, incluindo modelos de árvore de decisão, regressão linear, redes neuronais (NN), regressão de vetores de suporte (SVR) e agrupamento K-means (KMC). Os resultados do estudo indicaram que a Regressão de Vetores de Suporte (SVR) obteve a melhor pontuação de R^2 e o Erro Quadrático Médio (EQM).

Para desenvolver modelos de previsão de preços da Airbnb, Liu, (2021) utilizou uma combinação de análise de sentimento e classificação. O modelo de árvore de regressão obteve o menor Erro Médio Absoluto (MAE), quando a categorização de sentimentos e subjetividade foram usadas como características. Em contrapartida, o modelo Linear Lasso apresentou o erro médio absoluto mais elevado.

O estudo desenvolvido por (Gyódi, 2021) examinou o impacto da pandemia do COVID-19 nos setores hoteleiro e no Airbnb em nove cidades europeias, utilizando uma abordagem de regressão de dados de painel. A pesquisa analisou as distinções entre os dois modelos de negócio e como os anfitriões do Airbnb enfrentaram esta nova realidade. Além disso, utilizaram a análise de regressão de dados em painel para investigar o efeito da epidemia nos preços do Airbnb. Para estimar os preços, utilizou uma técnica de regressão Ridge, assim como o agrupamento em K-means. Além disso, o estudo explorou a utilização de regressão de vetores de suporte (SVR) juntamente com o *Gradient Boosting Tree Ensemble*. Além disso, o estudo avaliou o uso de Redes Neuronais (NN). O modelo mais eficaz que emergiu foi o SVR com kernel RBF.

3. Metodologia

3.1. Recolha e tratamento de dados

Neste estudo, escolheu-se Lisboa como local de investigação, o que alarga o âmbito geográfico de pesquisa para além dos estudos habitualmente realizados nas cidades dos Estados Unidos. Os dados foram obtidos através do InsideAirbnb. O InsideAirbnb baseia-se em informações publicamente disponíveis no *website* do Airbnb. Foram implementadas várias medidas para analisar, limpar e agregar os dados, a fim de facilitar o nosso estudo.

Os dados analisados foram extraídos a 13 e 14 de setembro de 2022. Este ficheiro contém 75 colunas, que incluem um conjunto de variáveis, sendo uma das variáveis o preço praticado por noite pelo anfitrião. Tem 19651 linhas e cada uma corresponde a um anúncio diferente. Todo o processo de limpeza e pré-processamento, análise exploratória de dados e modelação foram desenvolvidos com recurso ao software *Google Colab* e está disponível no GitHub https://github.com/diman25pt/Tese/tree/main. A limpeza e o pré-processamento de dados para este estudo envolveram várias etapas.

A fase de pré-processamento em qualquer atividade de ciência de dados é um dos componentes mais importantes. É fundamental, uma vez que estabelece os alicerces do modelo; se for efetuado incorretamente, pode ter um impacto substancial no modelo de previsão, tornando as conclusões redundantes. Este procedimento inclui a filtragem inicial dos dados, a limpeza, o tratamento de valores anómalos e de resultados em falta e a eliminação de duplicados.

Primeiramente, recolhemos o conjunto de dados de 13 e 14 de setembro de 2022, que inclui informações sobre todos os 19651 anúncios do Airbnb em Lisboa que estavam ativos na altura.

O segundo passo passou por analisar os valores únicos, nulos e tipo de dados de cada variável. Com base na tabela extraída (Anexo 1) podemos ver que as variáveis *bathrooms* e *calendar_update* não apresentam qualquer valor e a variável *scrape_id* tem apenas um valor único. Como tal terão de ser removidas do nosso *dataset*. Para além disso existem variáveis que apresentam apenas dois valores únicos, como é o caso de *last_scraped*, *source*, *host_is_superhost*, *host_has_profile_pic*, *host_identity_verified*, *has_availability*, *calendar_last_scraped e instant_bookable*. O problema deste tipo de variáveis é que caso tenham uma distribuição uniforme, isto é, grande parte dos dados corresponde a um dos valores, a variável torna-se irrelevante. Para analisar esta questão construímos um gráfico de barras para

cada uma das variáveis (Anexo 2). Como para todas estas variáveis um dos valores representava mais de 75% dos dados decidimos retirar todas as variáveis.

Do nosso conjunto de variáveis do *dataset* sabemos que as variáveis que são URL não irão ser utilizadas para previsão de preço. Deste modo, retirámos as seguintes colunas: *id, listing_url, scrape_id, picture_url, host_id, host_url, host_thumbnail_url* e *host_picture_url*.

Uma vez o nosso conjunto de dados é extenso decidimos retirar todas as colunas com mais de 15% de valores nulos. Às restantes variáveis com valores nulos apenas retirámos as respetivas linhas. Deste modo, retirámos as colunas: *neighborhood_overview, host_location, host_about, host_neighbourhood* e *neighbourhood*. Posto isto o nosso *dataset* apresentava 14312 linhas e 52 colunas.

Como não vamos dar uso de qualquer técnica de processamento de linguagem natural (NLP) podemos remover as colunas *name* e *description*.

De seguida, procedemos a uma análise mais detalhada entre variáveis. Como é o caso de colunas como, *minimum_nights, maximum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm*. Apesar de todas apresentarem informações diferentes as únicas que se demonstraram úteis para o nosso estudo foram *minimum_nights* e *maximum_nights*. As restantes foram eliminadas do nosso *dataset.*

Com base na visualização do *dataset,* identificámos que variáveis como *price, host_response_rate* e *host_acceptance_rate* eram *strings*, pelo que os símbolos de euro e percentagem foram removidos, e apenas a parte numérica foi mantida.

Quanto às variáveis relativas ao número total de anúncios de cada anfitrião, tais como *host_listings_count, host_total_listings_count, calculated_host_listings_count, calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms,* apenas mantivemos a coluna *calculated_host_listings_count*. Para além disso, de modo a fazermos uma análise geográfica vamos recorrer às variáveis *latitude, longitude* e *neighbourhood_group_cleansed*.

Posto isto, analisámos todas as variáveis de tipo objeto para que as mesmas pudessem ser admissíveis num modelo de previsão. Os atributos *host_verifications* e *license* foram excluídos, pois não têm qualquer utilidade para o estudo que pretendemos desenvolver. No entanto, *host_since, first_review* e *last_review* foram transformadas em dados do tipo data.

À semelhança de parâmetros analisados anteriormente, *host_response_time*, agrega a maioria dos seus dados em apenas um valor, neste caso, *'within an hour'*, e, portanto, será retirado do modelo. (Figura 2)

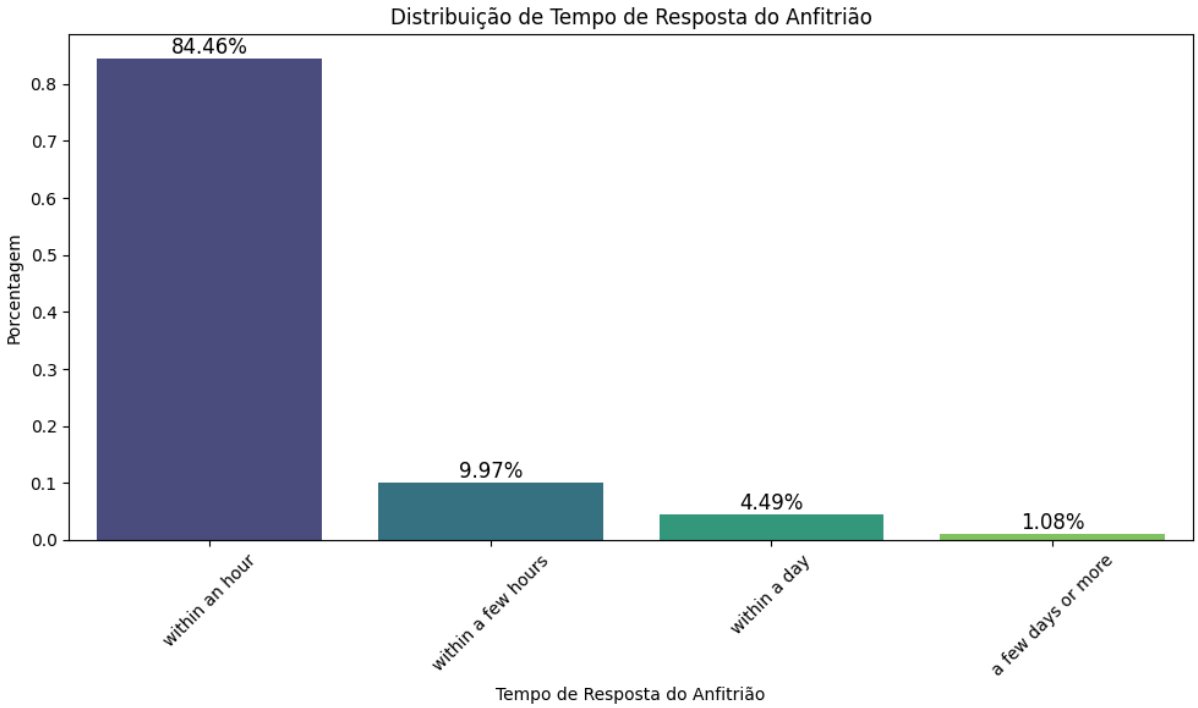


Figura 2 - Distrbuição da variável host_response_time

Ao verificar os dados de tipo booleano e categórico constatamos que *host_response_rate* e *host_acceptance_rate* estão muito agrupados no 1. O que torna estas duas variáveis irrelevantes.

Por conseguinte, passámos à análise de duas variáveis que em primeira instância parecem ser as mais relevantes para o nosso modelo, *property_type* e *room_type*. De modo a garantir uma correta recolha e tratamento dos nossos dados, foi necessário efetuar uma limpeza dos tipos de propriedades, uma vez que havia um grande número de categorias. Em vez de tentar agrupar tipos de propriedades semelhantes em grupos, vamos recorrer à variável *room_type* para distinguir o tipo de alojamento de cada anúncio.

Antes de nos debruçarmos sobre a nossa variável dependente, *price*, temos de fazer o devido tratamento das colunas *bathrooms_text* e *amenities*. Quanto às variáveis *bathrooms_text* apenas mantivemos o número e retirámos o dado quanto ao tipo de casa de banho, se partilhada ou privada. Sendo que *amenities* corresponde às comodidades oferecidas em cada anúncio, naturalmente apresenta um conjunto bastante diversificado de informação. Para uma leitura mais intuitiva e visual, construímos uma nuvem de palavras e um histograma com as palavras

mais frequentes. Como podemos ver na figura 3 e 4, identificámos as comodidades mais comuns como sendo *Kitchen*, *Wifi*, *Long term stays allowed*, *Essecials* e *Hangers*.

Figura 3- Nuvem de palavras da variável amenities

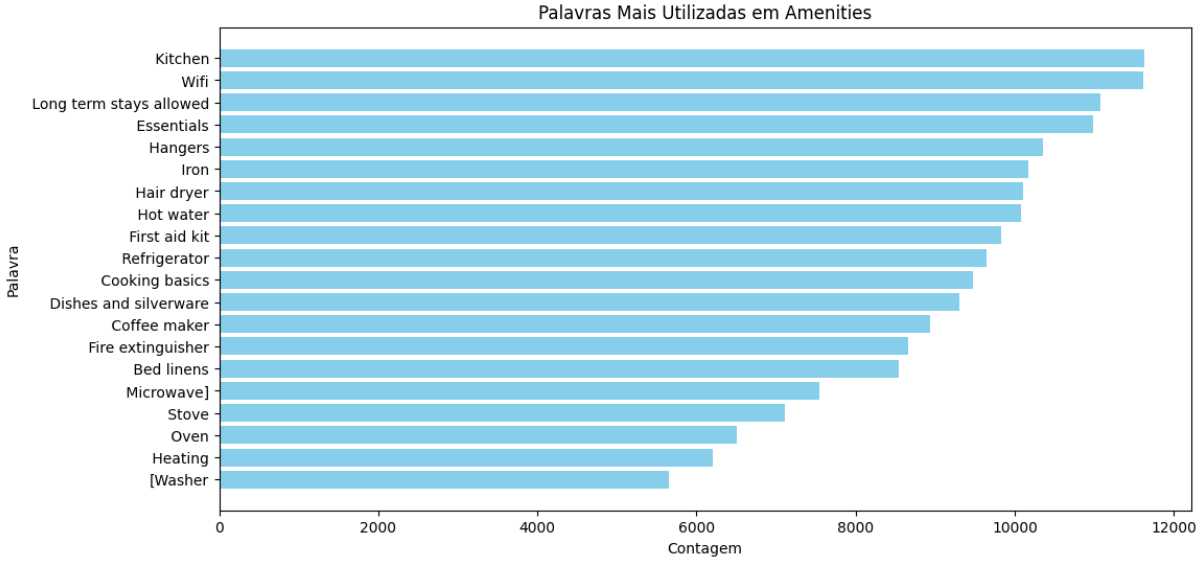


Figura 4- Gráfico de palavras mais frequentes na variável amenities

De todas as variáveis presentes no nosso *dataset* original, a variável *amenities* foi a que mais reflexão critica requereu. Por um lado, não faz sentido incorporar um *amenity* que quase nunca é utilizado. Deste modo não vamos considerar os que tem uma frequência inferior a 3000. Por outro lado, também não vamos admitir um *amenity* que se repita em praticamente todos os casos. E, portanto, tendo em conta estes fatores e com base em pesquisa e experiência pessoal

selecionámos os *amenities*, *Air conditioning*, *TV*, *Extra pillows and blankets* e *Heating*, como sendo os que mais impacto iriam ter no preço.

Por fim, com o objetivo de avaliar a nossa variável dependente, construímos um diagrama de caixa (figura 5) e uma tabela com as estatísticas descritivas (tabela 1) da mesma.

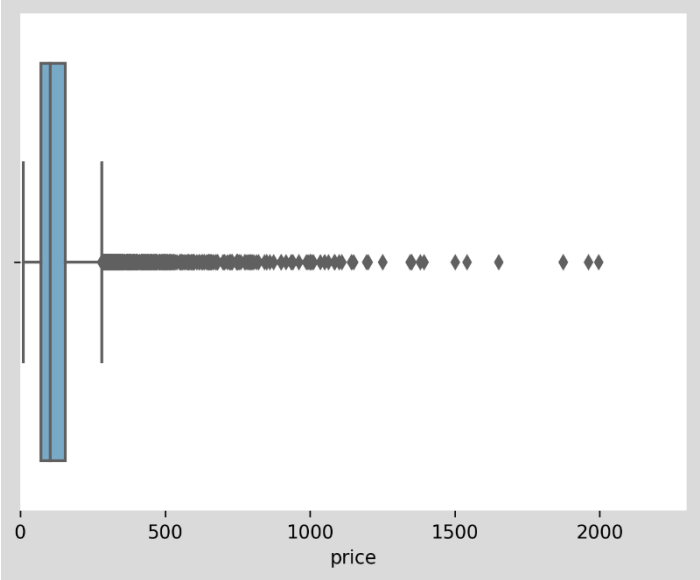


Figura 5 - Diagrama de caixa da variável price

Count	14312
Mean	134.456
Std	237.225
Min	10
25%	70
50%	102
75%	154
max	22000

Tabela 1 - Estatísticas descritivas da variável price

É difícil acreditar que o preço por noite de um Airbnb possa ir até 22000€. Em suma, existem muitos valores anómalos no preço que podem realmente alterar as nossas previsões. De modo a evitar que *outliers* influenciem de forma negativa o nosso modelo, restringimos o preço entre os 20€ e os 220€. Posto isto, como podemos ver na figura abaixo, a variável preço apresenta uma distribuição relativamente simétrica, com uma média de 103.14€ e uma mediana de 95€.

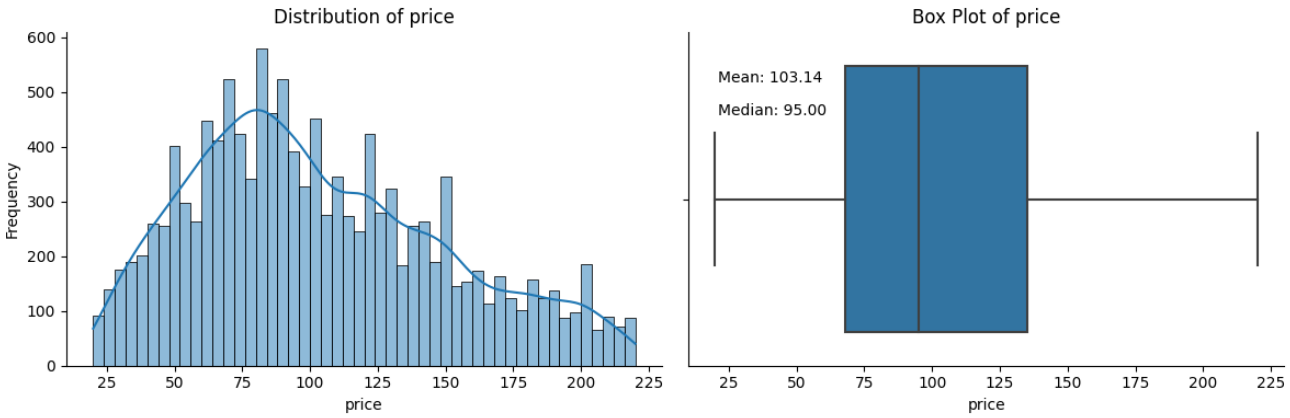


Figura 6 - Distribuição da variável price

Por último, desenhámos uma matriz de correlação para testar a multicolinearidade entre as variáveis. Todas as que apresentam um nível de correlação igual ou superior a 0.7, foram removidas.

Correlações altas, acima de 0,70 entre pares de variáveis indicam que as mesmas estão fortemente correlacionadas. A multicolinearidade é um problema que pode comprometer a interpretação dos resultados dos modelos.

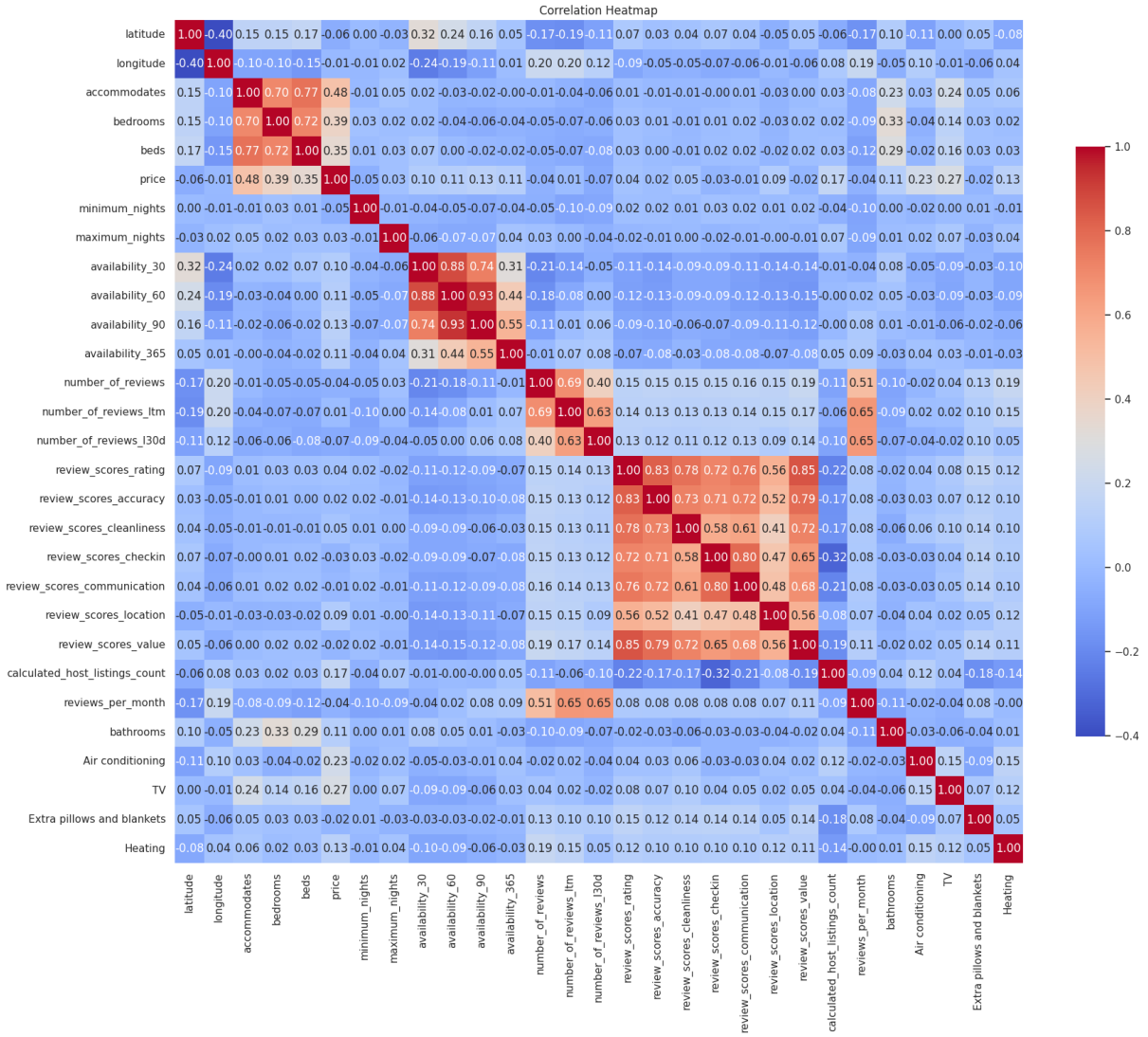


Figura 7 - Matriz de correlação original

Os parâmetros *beds*, *bedrooms* e *accommodates*. estão altamente correlacionados. Isto é compreensível uma vez que representam respetivamente o número de camas, o número de quartos e o número máximo de pessoas que acomodada. O número de pessoas acomodadas tem sido tradicionalmente um parâmetro de pesquisa de maior relevância nos estudos sobre o Airbnb, uma vez que é mais relevante para quartos privados e partilhados do que o número de

quartos ou número de camas. Como tal, mantemos *accommodates* em detrimento de *beds* e *bedrooms*.

Das três variáveis relativas à disponibilidade, *availability_30*, *availability_60* e *availability_90*, duas poderiam ser suprimidos, uma vez que estão altamente correlacionados entre si. Sendo que *availability_30* apresenta o nível de correlação mais baixo entre as três, excluímos do nosso dataset *availability_60* e *availability_90*.

Apesar de ser expectável que os *reviews* entre si sejam correlacionados, o que mais surpreendeu foi que a grande maioria dos anúncios incluídos no nosso dataset terem uma avaliação maioritariamente entre quatro e cinco. Como tal apenas iremos considerar o *review_scores_rating* e o *review_scores_location*.

Em suma, este foi o processo conduzido de modo a fazer a recolha e o tratamento dos nossos dados. A correlação entre as nossas variáveis foi testada uma última vez (figura 8) antes de iniciar a análise exploratória de dados.

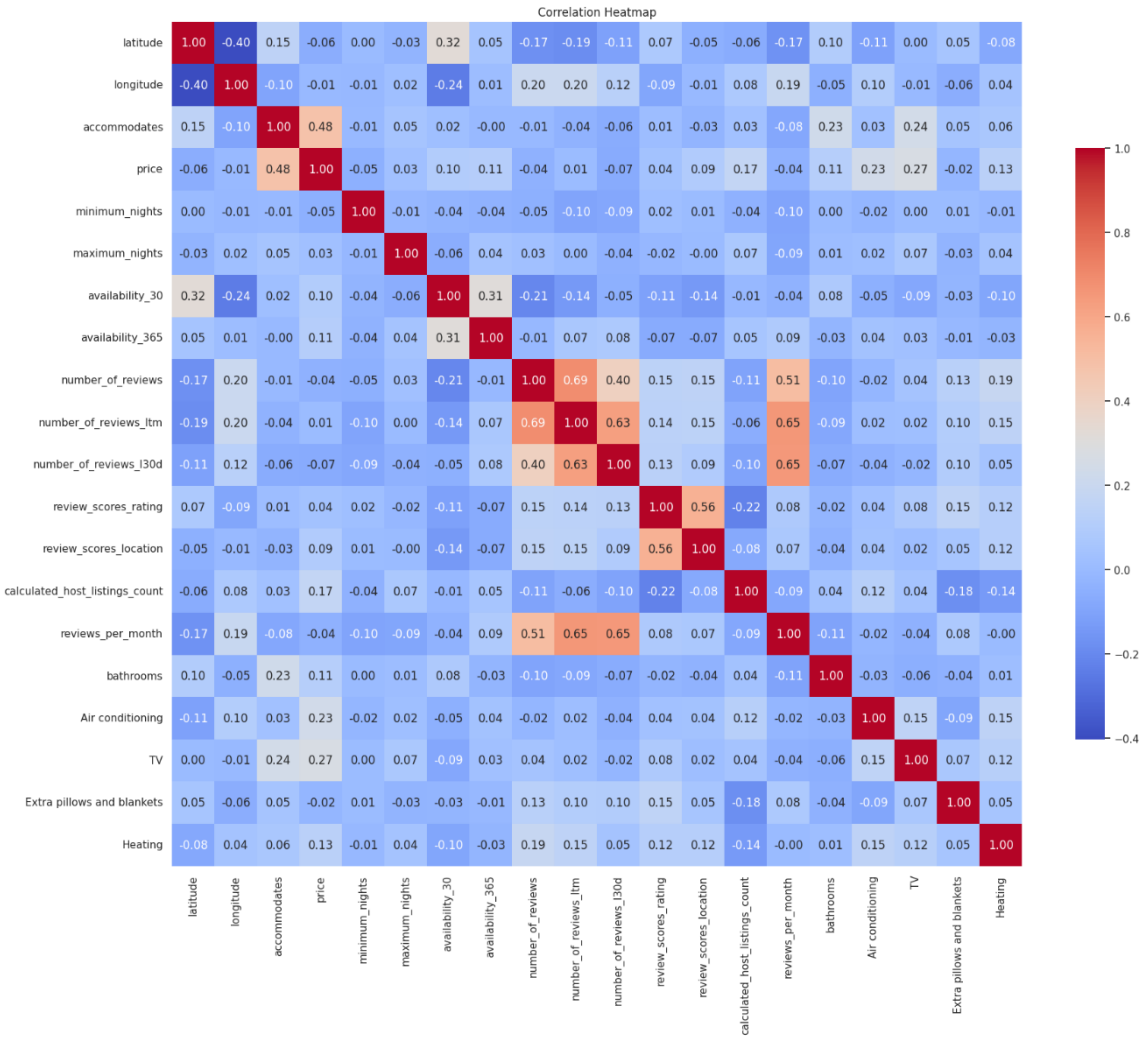


Figura 8 - Nova matriz de correlação das variáveis selecionadas

3.2. Análise exploratória de dados

Na segunda etapa do nosso projeto de investigação procedemos a uma análise exploratória de dados.

A Análise Exploratória de Dados (AED) é uma etapa primordial e fundamental no campo da ciência de dados, que procura compreender de forma aprofundada um conjunto de dados antes de se aplicarem técnicas estatísticas ou modelos de aprendizagem automática.

Primeiramente, analisámos as variáveis desenvolvidas no processo de recolha e tratamento de dados e com base nas variáveis *host_since*, *first_review* e *last_review*, construímos novas variáveis, *host_days_active*, *time_since_first_review* e *time_since_last_review*. Queremos perceber á quanto tempo o anfitrião utiliza a plataforma e o período que decorreu deste o primeiro e último comentário, respetivamente. Estas novas colunas têm como referência a data em que os dados foram recolhidos, 14 de setembro de 2022.

Com base na figura 9 percebemos que o intervalo do último comentário está concentrado entre as 0-2 semanas e entre 2-8 semanas. Por outro lado, temos um menor conjunto de anúncios em que o último comentário foi há mais de um ano. Esta situação levam-nos a questionar se este conjunto de anúncios está inativo, isto é, indisponível para reservar ou se simplesmente a relação preço e as condições oferecidas pelo anfitrião não correspondem corretamente.

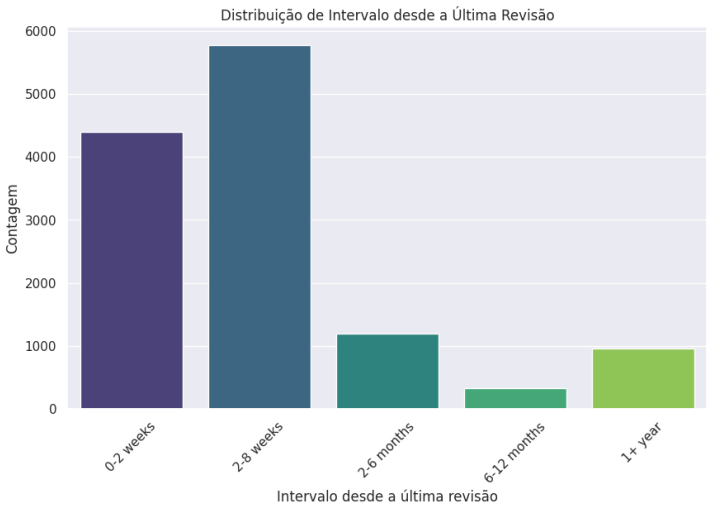


Figura 9 - Distribuição em intervalos da variável time_since_last_review

Por outro lado, apesar de não termos acesso ao período exato quando o anúncio foi publicado na plataforma do Airbnb, podemos aferir de forma parcial a evolução quanto à oferta na plataforma. Assumindo que o primeiro comentário para cada anúncio foi feito num período relativamente curto face à sua publicação, percebemos que a oferta tem aumentado de forma relativamente estável ao longo do tempo.

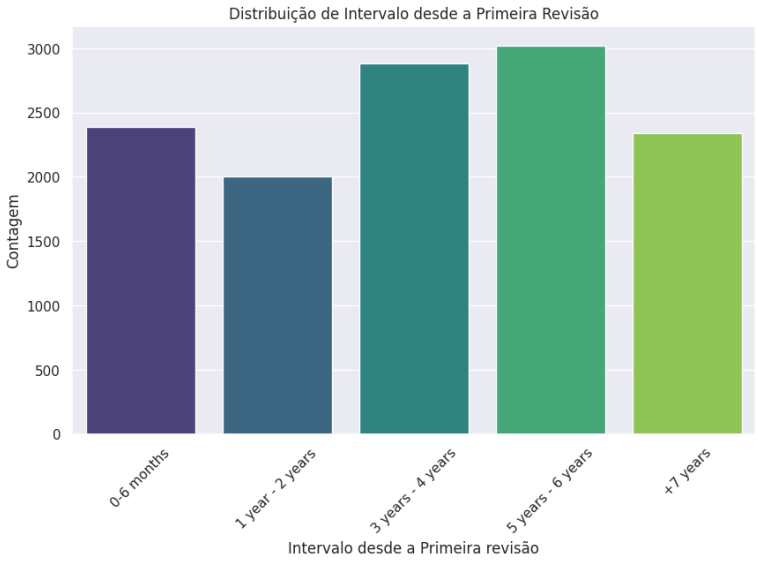


Figura 10 - Distribuição em intervalos da variável time_since_first_review

De seguida fomos avaliar o aspeto mais importante a observar, o preço. A variável preço pode ser vista de diferentes perspetivas. Em relação às mudanças de preço tendo em consideração a primeira avaliação de cada anúncio, verifica-se que o preço médio por noite de novos alojamentos e mais elevado que anúncios que já estão na plataforma há mais tempo (Anexo 3)

Não é de surpreender que as propriedades que acomodam mais pessoas atinjam preços notavelmente mais altos por noite (Figura 11).

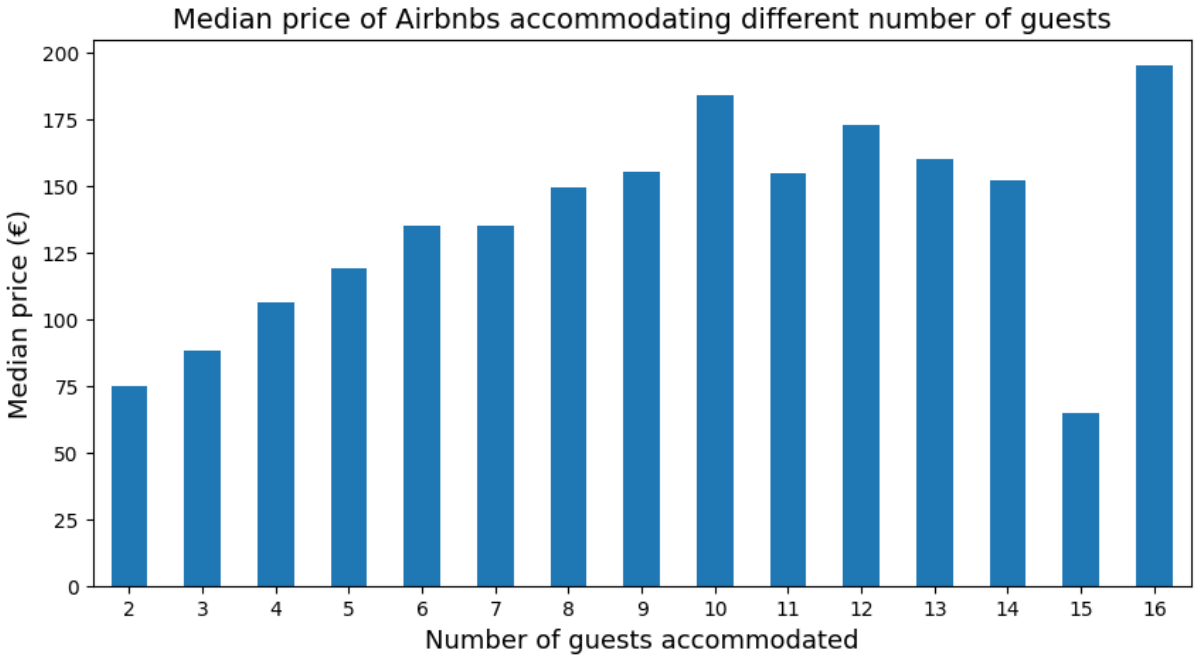


Figura 11 - Preço médio por noite face ao número máximo de pessoas que acomoda

Além disso, a representação dos preços num mapa mostra que os anúncios mais caros estão localizados perto da costa (Figura 12). Da mesma forma, os bairros mais caros são Vermelha e Painho e Figueiros com um preço médio de 166.33€ e 161.00€ por noite de alojamento, respetivamente. Os bairros mais económicos são Vila Franca de Xira e Ramalhal, com um preço médio de 45.86€ e 45.00 €, por esta ordem.

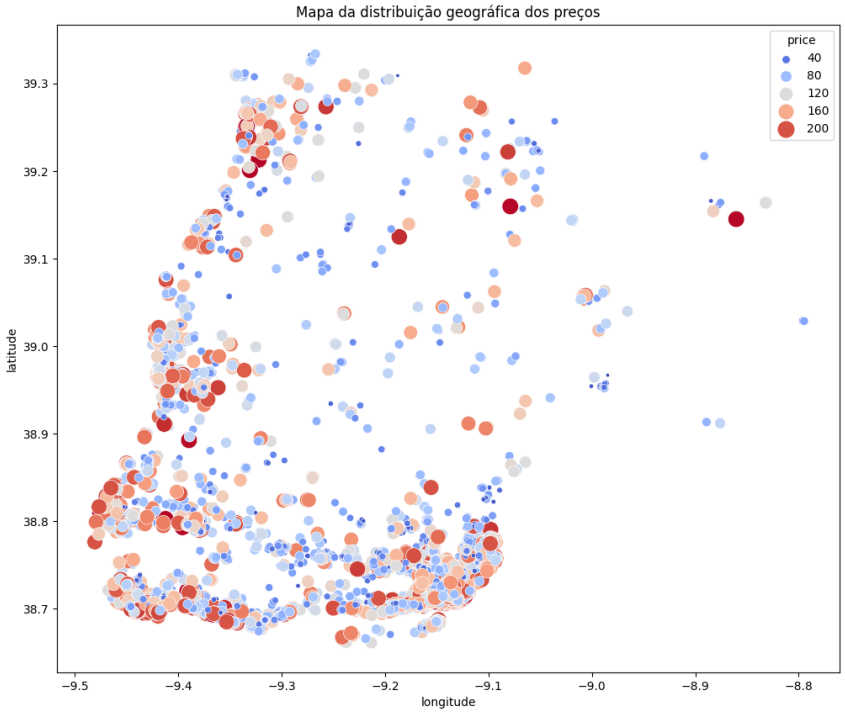


Figura 12 - Mapa da distribuição geográfica dos preços

Com base na figura 13, cerca de 77% dos anúncios são casas ou apartamentos completos (ou seja, está a alugar a propriedade completa). A maioria dos restantes são quartos privados (ou seja, está a alugar um quarto e possivelmente uma casa de banho, mas haverá outras pessoas na propriedade). Menos de 1% são quartos partilhados e por último e menos comum temos os quartos em hotel.

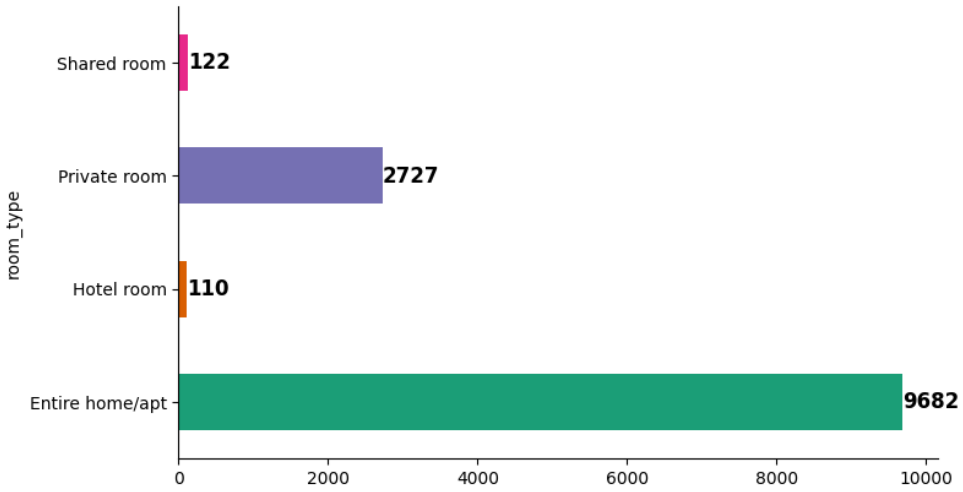


Figura 13- Distribuição da variável room_type

Apesar de ser evidente que existe mais oferta de alojamento junto ao litoral, decidimos averiguar se este fenómeno acontece independentemente do tipo de quarto. Com base nas figuras 14 a 17 conseguimos compreender que apesar de anúncios com tipo de quarto 'Private room' e 'Entire home/apt' terem uma oferta geográfica mais ampla, verifica-se uma concentração de oferta na costa, em especial na zona metropolitana de Lisboa.

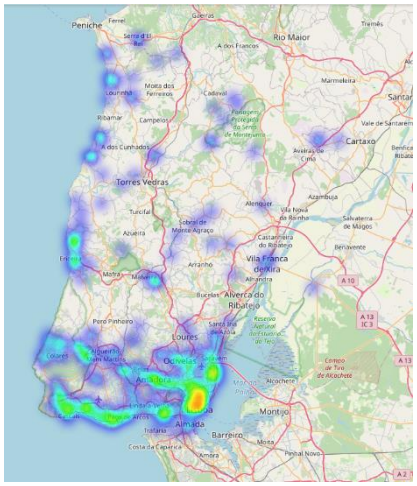


Figura 14 - Mapa de densidade - Private room

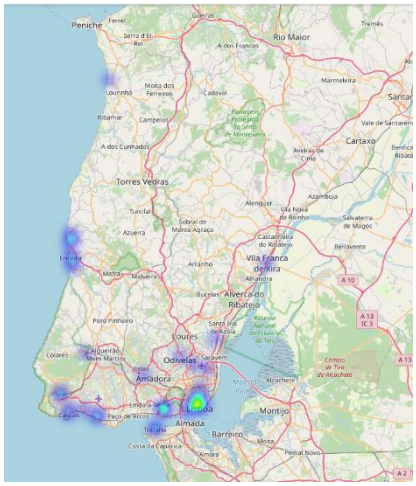


Figura 15 - Mapa de densidade - Shared room

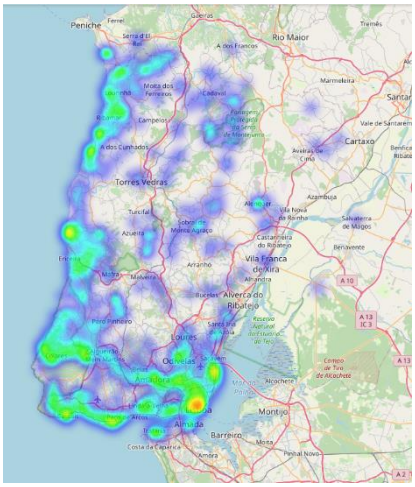


Figura 16 - Mapa de densidade - Entire home/apt

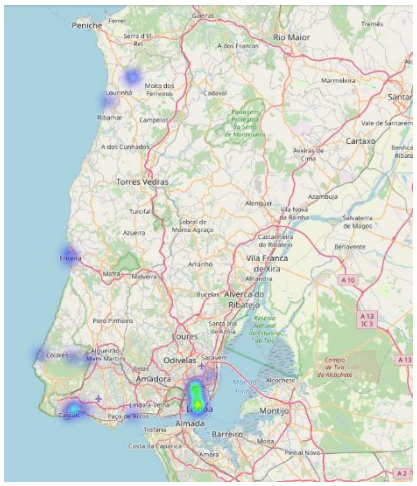


Figura 17 - Mapa de densidade - Hotel room

Antes de avançar para a modelação vamos ainda eliminar todas as colunas que utilizámos para ter uma perspetiva mais abrangente dos nossos dados, mas que não serão utilizados nos nossos modelos de previsão. Deste modo, retiramos do nosso *dataset* *first_review_interval*, *last_review_interval*, *first_review*, *last_review*, *neighbourhood_cleansed* e *host_since*. Quanto à variável categórica *room_type*, convertemos em quatro *dummies* distintas. Posto isto o nosso *dataset* apresentava 12641 linhas e 27 colunas, face aos 19651 registos iniciais e 74 colunas.

3.3. Modelação

Nesta última secção descreve-se a utilização de algoritmos de aprendizagem automática aos dados previamente tratados. Na fase de implementação de modelos estatísticos o nosso *dataset* era composto por 12641 linhas e 27 colunas. Na tabela abaixo é possível ver as variáveis e uma breve descrição das mesmas.

Colunas	Descrição
latitude	Latitude
longitude	Longitude
accommodates	Número máximo de pessoas na casa
price	Preço
minimum_nights	Mínimo de noites
maximum_nights	Máximo de noites
availability_30	Número de dias disponível nos próximos 30 dias
availability_365	Número de dias disponível por um ano
number_of_reviews	Número total de avaliações
number_of_reviews_ltm	Número total de avaliações no mês anterior
number_of_reviews_l30d	Número total de avaliações nos últimos 30dias
review_scores_rating	Avaliação geral do anúncio
review_scores_location	Avaliação da localização
calculated_host_listings_count	Número total de anúncios do anfitrião
reviews_per_month	Numero médio de avaliações por mês
bathrooms	Número de casas de banho
Air conditioning	Se existe ou não ar condicionado
TV	Se existe ou não TV
Extra pillows and blankets	Se existe ou não lençóis e cobertores extra
Heating	Se existe ou não aquecimento
time_since_first_review	Período desde a primeira avaliação
time_since_last_review	Período desde a última avaliação
host_days_active	Número de dias ativo
room_type_Entire home/apt	Tipo de quarto é Entire home/apt
room_type_Hotel room	Tipo de quarto é Hotel_room
room_type_Private room	Tipo de quarto é Private room
room_type_Shared room	Tipo de quarto é Shared room

Tabela 2 - Variáveis do dataset final

Uma vez que não existem valores nulos, uma vez que foram retirados na face de preparação os dados e o tipo de dados estão disponíveis para aplicar modelos, vamos proceder com o nosso trabalho de investigação. Para tal, começámos por construir um modelo de regressão OLS (*Ordinary Least Squares*) ou método dos mínimos quadrados.

Um modelo de regressão OLS (*Ordinary Least Squares*) é um método estatístico utilizado na análise de regressão para estimar a relação entre uma variável dependente (neste caso o preço) e uma ou mais variáveis independentes (todas as restantes variáveis do dataset final). Para poder atingir resultados melhores e ser possível comparar com trabalhos de investigação já desenvolvidos, aplicámos também modelos de *Machine Learning*, tais como, *Random Forest, XGBRegressor, Decision Tree e Extra Trees.* Para tal foi utilizada entre outras a biblioteca sklearn (Pedregosa et al., 2011) e XGBoost (Chen, & Guestrin, 2016).

```
import statsmodels.api as sm

# Adicione uma constante ao conjunto de dados
x = sm.add_constant(x)

# Ajuste o modelo OLS
model = sm.OLS(y, x).fit()

# Imprima um resumo do modelo
print(model.summary())
```

Figura 18 - Código fonte para construir um modelo OLS

O *Random Forest* (Ho, 1995) é um algoritmo de aprendizagem automática baseado em árvores de decisão, isto é, constrói várias árvores de decisão durante o período de treino e combina os resultados para obter uma previsão mais robusta e precisa. Cada árvore é treinada com uma amostra aleatória dos dados e as previsões são agregadas através de votação ou média, dependendo se é uma tarefa de classificação ou regressão. O Random Forest é conhecido pela sua capacidade de fornecer boas previsões num conjunto diverso de problemas.

XGBoost (Chen, & Guestrin, 2016) é um algoritmo de *machine learning* baseado em *gradient boosting.* É bastante utilizado no campo de ciência de dados devido ao seu desempenho excecional. O XGBRegressor é uma versão do XGBoost usada para tarefas de regressão. Cria uma série de árvores de decisão em sequência, onde cada nova árvore é treinada para corrigir os erros da previsão anterior. O XGBoost é conhecido pela sua eficiência, escalabilidade e capacidade de lidar com dados ausentes.

Uma árvore de decisão (*Decision Tree*) é um modelo que representa decisões e as respetivas consequências numa estrutura semelhante a uma árvore (Loh, 2011). Cada nó da árvore representa uma decisão ou um teste de uma variável, e os ramos da árvore representam os resultados possíveis. É a base para algoritmos mais complexos, como o Random Forest e o XGBoost.

Extra Trees, ou *Extremely Randomized Trees*, (Geurts, Ernst, & Wehenkel, 2006). é outra variação do algoritmo *Random Forest*. Assim como o *Random Forest*, o *Extra Trees* constrói várias árvores de decisão, mas com uma diferença fundamental. Enquanto o *Random Forest* procura o melhor corte em cada nó da árvore, o *Extra Trees* escolhe cortes aleatórios. O *Extra Trees* é conhecido por ser eficiente e útil para problemas de alta dimensionalidade.

Estes modelos são amplamente utilizados e escolhidos com base na natureza do problema, no tamanho dos dados e nas características específicas de cada algoritmo.

4. Resultados

Com base no modelo de regressão linear (*OLS*) e na descrição das variáveis independentes, é possível retirar algumas conclusões.

Quanto à localização, a latitude tem um impacto significativo no preço. Aumentos na latitude estão associados a uma diminuição no preço médio. Em termos geográficos permite perceber que quanto mais afastado da costa menor é o preço. No entanto, a longitude não parece ter um efeito significativo no preço.

Por outro lado, quanto maior a capacidade de acomodação, maior o preço médio. Para além disso, a quantidade mínima de noites e a quantidade máxima de noites têm impactos significativos no preço. Aumentos na quantidade mínima de noites estão associados a uma diminuição no preço, enquanto a quantidade máxima de noites tem um efeito menos pronunciado.

A disponibilidade nos próximos 30 dias e a disponibilidade ao longo do ano têm efeitos positivos no preço das listagens. Quanto mais disponível estiver uma propriedade, maior será o preço. De modo semelhante, avaliação e a localização afetam o preço. Avaliações mais altas estão associadas a preços mais altos. No entanto, o número de avaliações em si não parece ter um impacto significativo no preço.

Outro fator importante é o papel do anfitrião. O número de propriedades que um anfitrião gere também afeta positivamente os preços. Isso sugere que anfitriões com várias propriedades tendem a estabelecer preços mais elevados para as suas listagens.

A quantidade de casas de banho na propriedade é outro fator determinante. Propriedades com um maior número de casas de banho tendem a ser mais caras.

A presença de comodidades, como ar-condicionado e televisão, tem efeitos positivos nos preços das listagens. No entanto, a disponibilidade de travesseiros e cobertores adicionais está associada a preços mais baixos.

Por fim, diferentes tipos de quartos (como 'Apartamento Inteiro' ou 'Quarto partilhado') têm diferentes preços médios. Por exemplo, 'Apartamento Inteiro' tende a ser mais caro do que 'Quarto Compartilhado'.

Em geral, este modelo fornece informações úteis sobre os fatores que afetam o preço no Airbnb. Pode usado para prever os preços com base nas características das propriedades.

Para além do modelo de regressão linear, aplicámos modelos de *machine learning, nomeadamente, Random Forest, XGBRegressor, Decision Tree e Extra Trees*, os quais é possível observar os resultados estatísticos na tabela 3 e 4.

	R^2	MAE	RMSE
Random Forest	59.53%	21.93	29.32
XGBRegressor	62.25%	21.56	28.31
Decision Tree	14.49%	31.46	42.61
Extra Tress	59.91%	21.84	29.18

Tabela 3 - Resultados estatísticos do R^2 de treino e teste

	R^2 - Treino	R^2 - Teste
Random Forest	0.943	59.53%
XGBRegressor	0.904	62.25%
Decision Tree	1	14.49%
Extra Tress	1	59.91%

Tabela 4 - Resultados estatísticos dos modelos de machine learning

O modelo *Random Forest* tem um RMSE (*Root Mean Squared Error*) de 29.32, o que significa que, em média, as previsões têm um erro de cerca de 29.32 unidades monetárias. O MAE (Mean Absolute Error) é 21.93, indicando que as previsões têm um erro absoluto médio de 21.93 unidades monetárias. O R^2 (R-squared) é de 59.53%, o que sugere que este modelo explica 59.53% da variabilidade nos preços. Em termos simples, o modelo *Random Forest* tem um desempenho razoável na previsão de preços, mas ainda há erro não explicado.

O modelo *XGBoost* apresenta um desempenho melhor em comparação com o *Random Forest*. Contém um RMSE de 28.31 e um MAE de 21.56, o que indica previsões ligeiramente mais precisas. O R^2 é de 62.25%, sugerindo uma capacidade melhor de explicar a variabilidade nos preços. Isso indica que o *XGBoost* é um modelo mais robusto em relação ao *Random Forest* neste contexto.

O *Decision Tree*, por outro lado, apresenta um desempenho inferior aos outros dois modelos. Ele tem um RMSE significativamente mais alto de 42.61 e um MAE de 31.46, o que indica previsões menos precisas. O R^2 é de apenas 14.49%, o que significa que este modelo tem dificuldade em explicar a variabilidade nos preços.

Por fim, o modelo *Extra Trees* apresenta um desempenho muito semelhante ao *Random Forest*, com um RMSE de 29.18, um MAE de 21.84 e um R^2 de 59.91%.

Em resumo, o modelo *XGBoost* destaca-se como o melhor entre os quatro modelos testados, com o menor RMSE e MAE e o maior R^2. Isto sugere que o *XGBoost* é o mais preciso na previsão de preços com base nas variáveis disponíveis no *dataset* final.

No entanto, é importante lembrar que a interpretação desses resultados é específica para este conjunto de dados e pode não se aplicar a outras regiões ou contextos. Além disso, outros fatores não incluídos no modelo podem influenciar os preços das listagens. Portanto, essas conclusões devem ser consideradas como parte de uma análise mais ampla.

5. Discussão

A discussão entre os resultados obtidos e a revisão de literatura é essencial para contextualizar as conclusões deste estudo sobre a previsão de preços no Airbnb em Lisboa em relação às pesquisas existentes. Os resultados revelam a complexidade subjacente à dinâmica de preços neste mercado específico, identificando fatores de significativa importância que já foram corroborados pela literatura existente.

Um fator notável que se destaca é a relação positiva entre a capacidade de acomodação e os preços das listagens, o que está em consonância com a noção geral de que casas maiores e mais espaçosas tendem a ser mais caras. Além disso, a influência das políticas de reserva, especificamente a quantidade mínima e máxima de noites, também se mostra relevante na determinação dos preços, refletindo a prática comum de ajustar os preços com base na antecedência das reservas.

As avaliações dos hóspedes são outro fator crucial que se correlaciona com os preços. É amplamente aceite na literatura que avaliações mais positivas tendem a elevar os preços das listagens. Contudo, é intrigante notar que a quantidade de avaliações em si não parece ter um impacto significativo no preço, sugerindo que a qualidade das avaliações pode superar em importância a quantidade.

Anfitriões com múltiplas propriedades também são identificados como tendo um papel determinante nos preços, com uma tendência a fixar preços mais altos, alinhando-se com pesquisas que destacam a estratégia de anfitriões profissionais em relação à definição de preços.

A variação nos preços médios entre diferentes tipos de quartos, como 'Apartamento Inteiro' ou 'Quarto Compartilhado', está em consonância com a literatura existente que aponta para diferenças de preços com base na privacidade e nas comodidades oferecidas.

Por último, a comparação dos modelos de previsão enfatiza o *XGBoost* como a técnica mais precisa para a previsão de preços. Este resultado é consistente com estudos anteriores que também destacaram o *XGBoost* como uma abordagem eficaz na previsão de preços no Airbnb.

No entanto, não basta restringe a nossa discussão à mera análise da metodologia na forma de construção dos modelos, sendo igualmente imperativo examinar os resultados de precisão. Nesse sentido, a tabela subsequente apresenta uma compilação dos modelos que obtiveram os melhores desempenhos em diversos estudos.

Autor	Técnicas Utilizadas	Nível de Precisão (R^2)
Luo et al. (2019)	Rede Neuronal	0.769
Rezazadeh Kalehbasti et al. (2019)	Vetores de Suporte	0.69
Liu (2021)	Análise de Sentimento	0.481

Tabela 5 - Compilação dos modelos com os melhores desempenhos

Analisando tanto o nosso trabalho quanto aos resultados obtidos por outros autores, é evidente que a precisão dos modelos de previsão varia substancialmente. O coeficiente de determinação, representado como R^2, é uma métrica fundamental para avaliar o desempenho dos modelos de previsão, uma vez que quantifica a quantidade de variação nos preços dos anúncios que os modelos são capazes de explicar. Portanto, quanto maior o valor de R^2, maior a precisão do modelo. Confrontando esses resultados com os de outros estudos, como o de Luo et al. (2019), que obteve um R^2 de 0.769 com Redes Neuronal (NN), o nosso modelo XGBRegressor não supera essa precisão. No entanto, é essencial lembrar que a comparação de métricas de precisão entre estudos pode ser complexa, uma vez que os conjuntos de dados, as características e os contextos variam. Portanto, embora o nosso modelo tenha demonstrado um desempenho sólido, é importante reconhecer que a eficácia de um modelo pode ser altamente específica para o contexto em que foi desenvolvido. Além disso, os modelos de previsão têm sempre espaço para melhoria. Como tal, os nossos resultados podem servir como base para futuras investigações.

6. Conclusão, limitações e trabalhos futuros

O Airbnb é uma plataforma online que permite que as pessoas encontrem alojamento, como casas, apartamentos, quartos individuais e até mesmo espaços únicos, como castelos e casas na árvore, para estadias curtas ou prolongadas.

A dinâmica neste setor é influenciada por uma série de fatores, tornando a previsão de preços uma tarefa desafiadora, mas crucial para os anfitriões e para os hóspedes. Compreender os determinantes dos preços e desenvolver modelos de previsão precisos é de grande importância para todos os envolvidos neste ecossistema. Neste contexto, a análise de dados e a utilização de algoritmos de aprendizagem automática são ferramentas valiosas para tomar decisões informadas sobre a previsão de preços. Neste projeto de investigação, foram utilizados diferentes algoritmos de aprendizagem automática. O modelo que apresentou o resultado mais consistente foi o XGBRegressor, sendo capaz de explicar 62% da variação dos preços.

É crucial salientar que a nossa investigação apresenta restrições, como é inerente a qualquer estudo. Em primeiro lugar, considerou apenas o preço de cada anúncio num momento específico, o que não permite ter em consideração as variações sazonais na importância de certos atributos ou nas flutuações de preços sazonais. Durante o curso deste estudo, deparamo-nos com diversas limitações relacionadas com os dados. Por um lado, algumas variáveis que possivelmente seriam relevantes continham uma alta quantidade de valores ausentes, o que inviabilizou a sua inclusão na análise. Por outro lado, embora o nosso conjunto de dados inicial fosse substancial em termos de dados, faltaram-nos informações sobre variáveis que são frequentemente utilizadas em outros estudos, como as taxas aplicadas tanto ao anfitrião quanto ao hóspede no momento da reserva. Adicionalmente, enfrentámos limitações em relação ao software utilizado, o *Google Colab*, que, embora seja uma ferramenta valiosa, apresentava restrições de processamento na versão gratuita.

Para melhorias futuras, mais atributos deveriam ser considerados como por exemplo avaliar o impacto de eventos globais, como foi o caso da pandemia COVID-19, e locais, tais como eventos específicos em Lisboa que afetem a disponibilidade e os preços dos anúncios. Seria igualmente interessante analisar como a sazonalidade impacta a previsão de preços. Para além disto, incorporar dados exógenos como é o caso de novas regulamentações no mercado de alojamento. Por último, apesar de cada país e região ter a sua realidade especifica comparar Lisboa com outras cidades, especialmente europeias, poderia enriquecer a literatura neste campo de investigação.

A investigação confirma muitas das tendências identificadas na revisão da literatura e destaca a importância de determinadas variáveis para uma correta previsão de preços. Os resultados da investigação também apontam para a eficácia dos algoritmos de aprendizagem automática na previsão.

Referências

Aparicio, J. T., Aparicio, M., & Costa, C. J. (2023). Design Science in Information Systems and Computing. In Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (pp. 409-419). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9331-2_35

Aparicio, S., Aparicio, J. T., & Costa, C. J. (2019). Data Science and AI: trends analysis. In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI.2019.8760820

Arriaga, A., & Costa, C. J. (2023). Modeling and Predicting Daily COVID-19 (SARS-CoV-2) Mortality in Portugal: The Impact of the Daily Cases, Vaccination, and Daily Temperatures. In Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (pp. 275-285). Singapore: Springer Nature Singapore.

Aydin, R. (2019). "How 3 guys turned renting air mattresses in their apartment into a $31 billion company, Airbnb". Business Insider. https://www.businessinsider.com/how-airbnb-was-founded-a-visual-history-2016-2. Consultado October 14, 2023.

Bao, H., & Shah, S. (2020). The Impact of Home Sharing on Residential Real Estate Markets. *Journal of Risk and Financial Management*, 13, 161. https://doi.org/10.3390/jrfm13080161

Cassell, M. K., & Deutsch, A. M. (2020). Urban challenges and the gig economy: How German cities cope with the rise of Airbnb. German Politics, 1-22. https://doi.org/10.1080/09644008.2020.1719072

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. CRISP-DM Consortium

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI49556.2020.9140932.

Costa, C. J., & Aparicio, J. T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In *Advances in Parallel & Distributed Processing, and Applications: Proceedings from PDPTA'20, CSC'20, MSV'20, and GCC'20* (pp. 65-75). Springer International Publishing. https://doi.org/10.1007/978-3-030-69984-0_7

Costa, C. J., & Aparicio, M. (2023). Applications of data science and artificial intelligence. Applied Sciences, 13(15), 9015.

Custódio, J. P. G., Costa, C. J., & Carvalho, J. P. (2020, June). Success prediction of leads–A machine learning approach. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Garcia-López, M. À., Jofre-Monseny, J., Martínez-Mazza, R., & Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, *119*, 103278. https://doi.org/10.1016/j.jue.2020.103278

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3-42. https://doi.org/10.1007/s10994-006-6226-1

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L., & Morton, J. (2018). Use of dynamic pricing strategies by Airbnb hosts. International Journal of Contemporary Hospitality Management, 30(1), 2-20.

Gyódi, K. (2021). Airbnb and hotels during COVID-19: Different strategies to survive. International *Journal of Culture, Tourism, and Hospitality Research*. https://doi.org/10.1108/IJCTHR-09-2020-0221

Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE. https://doi /10.1109/ICDAR.1995.598994

Kwok, L., & Xie, K. L. (2019). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros?. *International Journal of Hospitality Management*, 82, 252-259.

Liu, P. (2021). Airbnb Price Prediction with Sentiment Classification. Master's Projects. https://doi.org/10.31979/etd.cfxc-m67z

Loh, W. Y. (2011). Classification and regression trees. *WIRE: data mining and knowledge discovery*, 1(1), 14-23. https://doi.org/10.1002/widm.8

Luo, Y., Zhou, X., & Zhou, Y. (2019). *Predicting Airbnb Listing Price Across Different Cities*;. Stanford University: Stanford, CA, USA.

Mergulhao, M., Palma, M., & Costa, C. J. (2022). A Machine Learning approach for shared bicycle demand forecasting. In 2022 17th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE https://doi.org/10.23919/cisti54924.2022.9820507

d'Orey Pape, R., Costa, C. J., Aparicio, M., & de Castro Neto, M. (2023). Determinants of City Mobile Applications Usage and Success. In Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (pp. 605-613). Singapore: Springer Nature Singapore https://doi.org/10.1007/978-981-19-9331-2_52

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.

Polisetty, A., & Kurian, J. S. (2021). The future of shared economy: A case study on Airbnb. *FIIB Business Review*, 10(3), 205-214. https://doi.org/10.1177/23197145211003504

Rezazadeh Kalehbasti, P., Nikolenko, L., Rezaei, H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds) *Machine Learning and Knowledge Extraction. CD-MAKE 2021*. Lecture Notes in Computer Science, vol 12844. Springer, Cham. https://doi.org/10.1007/978-3-030-84060-0_11

Samadani, S., & Costa, C. J. (2021). Forecasting real estate prices in Portugal: A data science approach. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI52073.2021.9476447.

Tavares, V. C., d'Água, J., Mendes, G., Peso, E., & Costa, C. J. (2022). Predicting the Portuguese GDP Using Three Different Computational Techniques. In World Conference on Information Systems and Technologies (pp. 513-523). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-04829-6_46

Thaichon, P., Surachartkumtonkun, J., Singhal, A., & Alabastro, A. (2020). Host and guest value co-creation and satisfaction in a shared economy: The case of Airbnb. *Journal of Global Scholars of Marketing Science*, 30(4), 407-422.

Xie, K. L., & Kwok, L. (2017). The effects of Airbnb's price positioning on hotel performance. International Journal of Hospitality Management, 67, 174–184. https://doi.org/10.1016/j.ijhm.2017.08.011

Yrigoy, I. (2016). The impact of Airbnb in the urban arena: towards a tourism-led gentrification? The case-study of Palma old quarter (Mallorca, Spain). *Turismo y crisis, turismo colaborativo y ecoturismo. XV Coloquio de Geografía del Turismo*, el Ocio y la Recreación de la AGE, 281-289.

Zekanovic-Korona, L., & Grzunov, J. (2014). Evaluation of shared digital economy adoption: Case of Airbnb. *In 2014 37th International Convention on Information and*

Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1574-1579). IEEE.

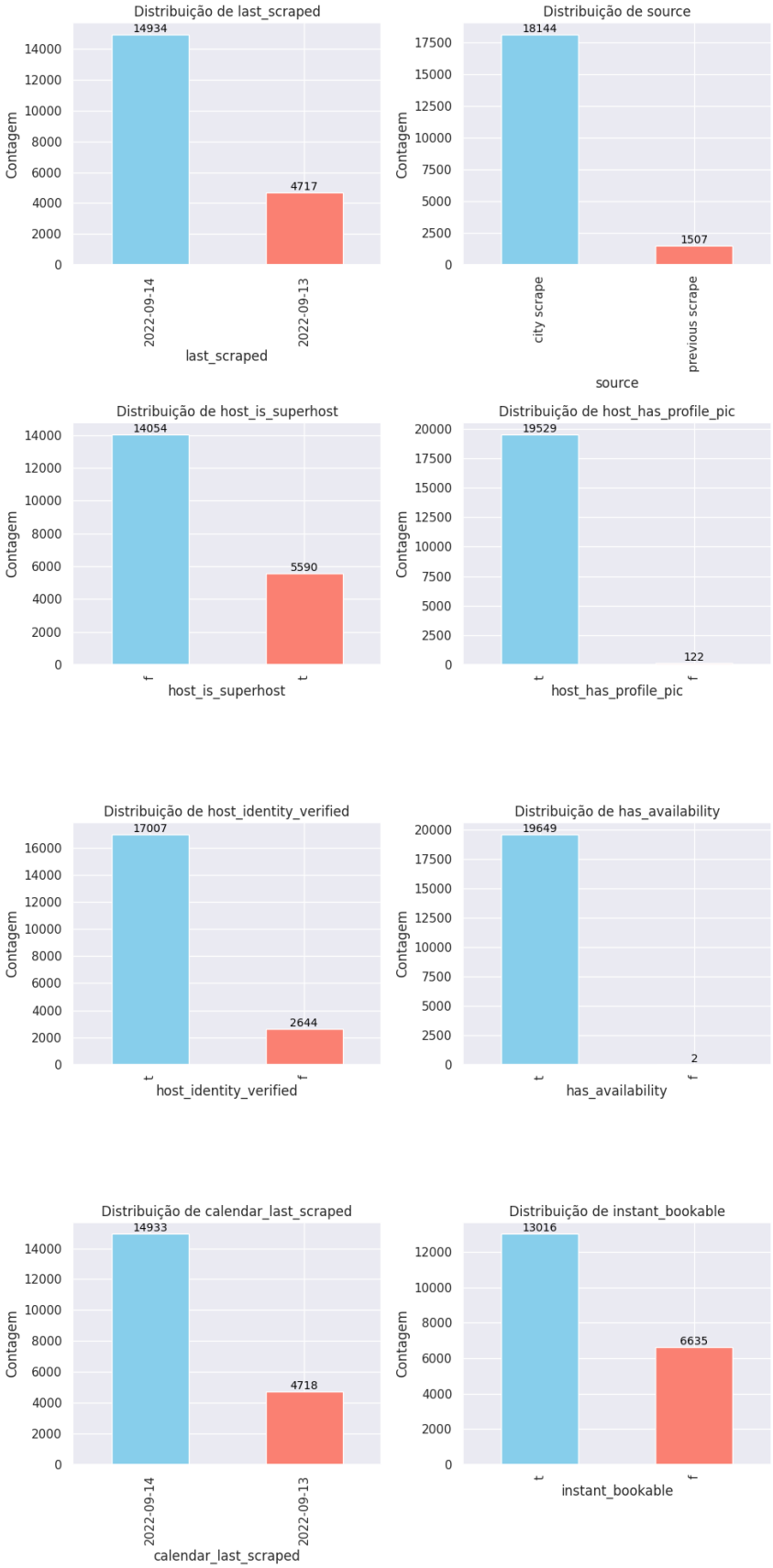
Zhang, Z., Chen, R. J., Han, L. D., & Yang, L. (2017). Key factors affecting the price of Airbnb listings: A geographically weighted approach. *Sustainability*, 9(9), 1635. https://doi.org/10.3390/su9091635

Anexos

Coluna	Valores ausentes	Valores Únicos
id	0.000000	19651
listing_url	0.000000	19651
scrape_id	0.000000	1
last_scraped	0.000000	2
source	0.000000	2
name	0.040710	19321
description	0.096687	18359
neighborhood_overview	36.257697	9415
picture_url	0.000000	19386
host_id	0.000000	8063
host_url	0.000000	8063
host_name	0.000000	3164
host_since	0.000000	3288
host_location	21.479823	539
host_about	40.690041	3990
host_response_time	9.241260	4
host_response_rate	9.241260	65
host_acceptance_rate	6.895323	93
host_is_superhost	0.035622	2
host_thumbnail_url	0.000000	7978
host_picture_url	0.000000	7978
host_neighbourhood	51.040660	131
host_listings_count	0.000000	88
host_total_listings_count	0.000000	102
host_verifications	0.000000	8
host_has_profile_pic	0.000000	2
host_identity_verified	0.000000	2
neighbourhood	36.257697	482
neighbourhood_cleansed	0.000000	128
neighbourhood_group_cleansed	0.000000	16
latitude	0.000000	10151
longitude	0.000000	11433
property_type	0.000000	87
room_type	0.000000	4
accommodates	0.000000	17
bathrooms	100.000000	0
bathrooms_text	0.188286	47
bedrooms	4.549387	16
beds	0.987227	26
amenities	0.000000	18218
price	0.000000	693
minimum_nights	0.000000	60
maximum_nights	0.000000	164
minimum_minimum_nights	0.000000	59

maximum_minimum_nights	0.000000	67
minimum_maximum_nights	0.000000	145
maximum_maximum_nights	0.000000	153
minimum_nights_avg_ntm	0.000000	241
maximum_nights_avg_ntm	0.000000	1557
calendar_updated	100.000000	0
has_availability	0.000000	2
availability_30	0.000000	31
availability_60	0.000000	61
availability_90	0.000000	91
availability_365	0.000000	366
calendar_last_scraped	0.000000	2
number_of_reviews	0.000000	484
number_of_reviews_ltm	0.000000	134
number_of_reviews_l30d	0.000000	23
first_review	9.841738	3139
last_review	9.841738	1244
review_scores_rating	9.841738	178
review_scores_accuracy	10.060557	158
review_scores_cleanliness	10.060557	186
review_scores_checkin	10.060557	167
review_scores_communication	10.060557	161
review_scores_location	10.060557	154
review_scores_value	10.055468	174
license	6.930945	12600
instant_bookable	0.000000	2
calculated_host_listings_count	0.000000	54
calculated_host_listings_count_entire_homes	0.000000	51
calculated_host_listings_count_private_rooms	0.000000	28
calculated_host_listings_count_shared_rooms	0.000000	11
reviews_per_month	9.841738	726

Anexo 1 - DataFrame com valores nulos e valores únicos



Anexo 2 - Distribuição de variáveis com apenas dois valores únicos

Anexo 3 - Preço médio por noite face ao ano da primeira avaliação