Relatório Analítico

Karl Richard Busse Filho

Relatório Analítico

Karl Richard Busse Filho

São Paulo

2024

Sumário das seções

1. Introdução

Objetivo

Este relatório se propõe a analisar os dados coletados e pré-processados de acomodações registradas no Airbnb, com o objetivo de encontrar relações entre as diferentes características destas acomodações com o valor da diária.

Por fim, o objetivo final é preparar o caminho para futuras análises e pesquisas. Quanto mais conhecermos o mercado, melhor será o posicionamento do cliente.

Descrição do Dataset

O dataset está disponível para consulta e download no link abaixo:
https://github.com/karlfilho/Airbnb-Price-Prediction/blob/main/final_dataset.csv

Ele foi obtido através do *scraping* do Airbnb, em duas etapas, e depois passou por pré-processamento para limpeza e organização dos dados até sua forma atual. Primeiro pesquisamos por acomodações em Campos do Jordão, para 2 adultos, no período entre 17 e 19 de Setembro. Esse padrão de busca cobre o local de interesse, com o tipo de hóspede mais comum, e num período de baixa temporada, em que a maioria das acomodações está livre e disponível para aparecer na busca. O Airbnb não informa exatamente quantas acomodações estão disponíveis, mas diz que foram mais de 1000 no momento da busca.

Fizemos o scraping em duas fases. Inicialmente coletamos o URL de todas as páginas das acomodações listadas na busca, e geramos o primeiro dataset, chamado “airbnb_frontPage.json” (disponível no repositório Github). Em seguida montamos uma lista com essas URLs e usamos novamente o software para iterar os endereços e coletar os dados, gerando o segundo dataset, chamado “airbnb_rooms.json” (também disponível no repositório Github). Foram coletados 670 registros. A próxima etapa foi a união dos dois datasets.

Após a união dos datasets fizemos a limpeza dos dados. Nesta fase identificamos 32 acomodações sem avaliações e há pouco tempo listadas no Airbnb, então preenchemos esses valores com ‘0’ (zero) e criamos uma nova coluna para identificar se a acomodação é nova ou não. Também identificamos 27 acomodações com preços errados (< R$80,00 a diária), então retiramos essas acomodações. Por fim, agregamos as centenas de *amenities* em grandes grupos e separamos em colunas no estilo one-hot encoding.

O dataset possui 643 registros e 46 colunas, abrangendo diversos aspectos das acomodações e suas características. Abaixo está uma descrição detalhada das colunas:

- `roomType`: Tipo de acomodação (ex.: Loft, Chalé, Quarto).
- `roomPrice`: Preço da diária da acomodação em Reais (R$).
- `roomURL`: URL da página da acomodação no Airbnb.
- `qualityBadge`: Selo de qualidade (ex.: preferido, superhost).
- `rating`: Avaliação média da acomodação (0-5 estrelas).
- `countReviews`: Número de avaliações recebidas pela acomodação.
- `Air Conditioning`: Indica se a acomodação possui ar condicionado (0 ou 1).
- `TV`: Indica se a acomodação possui TV (0 ou 1).
- `Hair Dryer`: Indica se a acomodação possui secador de cabelo (0 ou 1).
- `Bathroom`: Indica se a acomodação possui banheiro (0 ou 1).
- `Ethernet connection`: Indica se a acomodação possui conexão Ethernet (0 ou 1).
- `Kitchen`: Indica se a acomodação possui cozinha (0 ou 1).
- `Elevator`: Indica se a acomodação possui elevador (0 ou 1).
- `Luggage Dropoff Allowed`: Indica se a acomodação permite depósito de bagagens (0 ou 1).
- `Smoke Alarm`: Indica se a acomodação possui alarme de fumaça (0 ou 1).
- `WiFi`: Indica se a acomodação possui WiFi (0 ou 1).
- `Parking`: Indica se a acomodação possui estacionamento (0 ou 1).
- `Pets Allowed`: Indica se a acomodação permite animais de estimação (0 ou 1).
- `EV Charger`: Indica se a acomodação possui carregador para veículos elétricos
- `Bedroom`: Indica se a acomodação possui quarto (0 ou 1).
- `Fire pit`: Indica se a acomodação possui lareira (0 ou 1).
- `Lit path to the guest entrance`: Indica se há caminho iluminado para a entrada da acomodação (0 ou 1).
- `Waterfront`: Indica se a acomodação está à beira de um corpo de água (0 ou 1).
- `Long term stays allowed`: Indica se a acomodação permite estadias de longo prazo (0 ou 1).
- `Bathtub`: Indica se a acomodação possui banheira (0 ou 1).
- `Laundry room`: Indica se a acomodação possui lavanderia (0 ou 1).
- `Security Cameras`: Indica se a acomodação possui câmeras de segurança (0 ou 1).
- `Baby bathtub`: Indica se a acomodação possui banheira para bebês (0 ou 1).
- `Stove`: Indica se a acomodação possui fogão (0 ou 1).
- `Refrigerator`: Indica se a acomodação possui geladeira (0 ou 1).
- `Smoking allowed`: Indica se a acomodação permite fumar (0 ou 1).

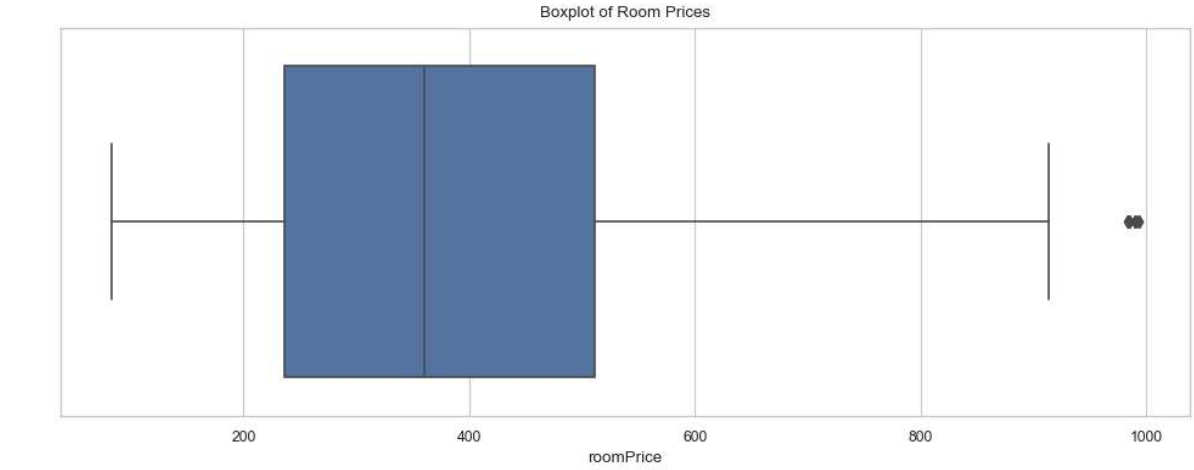
- `Patio`: Indica se a acomodação possui pátio (0 ou 1).
- `High Chair`: Indica se a acomodação possui cadeira de alimentação infantil (0 ou 1).
- `Sauna`: Indica se a acomodação possui sauna (0 ou 1).
- `Crib`: Indica se a acomodação possui berço (0 ou 1).
- `Washer`: Indica se a acomodação possui máquina de lavar (0 ou 1).
- `Accessible`: Indica se a acomodação é acessível para pessoas com deficiência, com rampas, corrimões, etc (0 ou 1).
- `Breakfast`: Indica se a acomodação oferece café da manhã (0 ou 1).
- `is_new`: Indica se a acomodação é nova (0 ou 1).

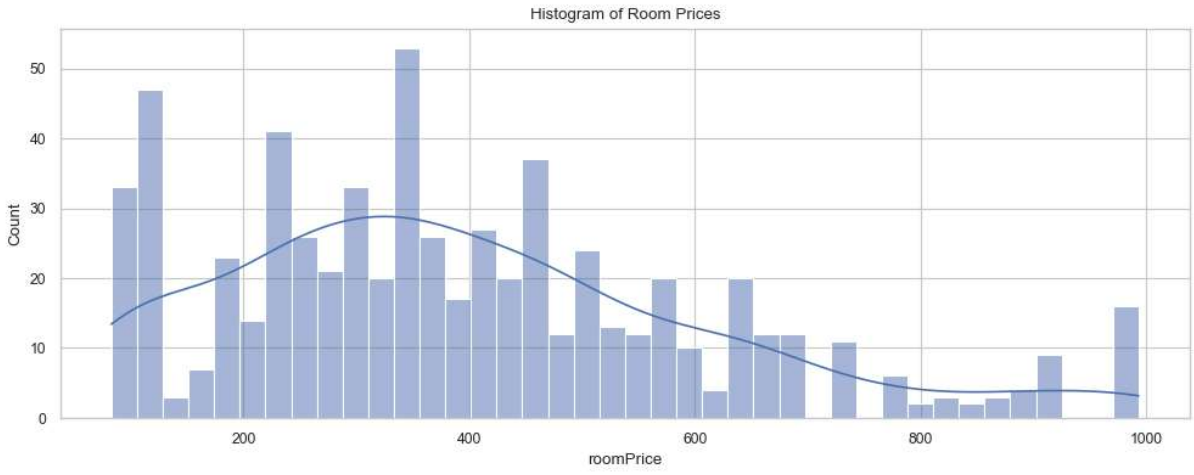
2. Análise descritiva

Para esta primeira parte foi analisado o conjunto geral dos dados. Posteriormente, a pedido do cliente, providenciaremos uma análise mais específica dos concorrentes diretos.

Preço (R$)

Média	Mediana	Mínimo	Máximo	Desvio-padrão
397,14	360,00	83,00	993,00	215,93





Reparem que o gráfico boxplot mostra que há pelo menos 3 outliers, ou seja, com preços muito acima do geralmente praticado. Não iremos removê-los, pois além de serem raros, podem trazer informações importantes. O histograma também mostra esses outliers perto dos R$1000,00 pela diária.

Os preços não seguem uma distribuição gaussiana, com assimetria à direita, o que significa que existem acomodações com preços elevados que estão aumentando a média. A maioria dos preços está concentrada entre R$150 e R$600 a diária. Observa-se um pico significativo entre R$100 e R$400, indicando que muitas acomodações são oferecidas nessa faixa de preço. O preço mais frequente é R$100, com mais de 60 acomodações listadas.

Separamos as acomodações em três grupos:

- **Baixo preço**: até o percentil 25
- **Médio preço**: entre o percentil 25 e o percentil 75
- **Alto preço**: acima do percentil 75

Escolhemos dividir pelo percentil e não pelo z-score (múltiplos do desvio-padrão) por causa da assimetria da distribuição e pela presença de outliers. Dessa maneira, as estatísticas para cada grupo são:

Grupo	Média	Mediana	Mínimo	Máximo	DP
Baixo preço	154,30	149,00	83,00	236,00	55,14
Médio preço	368,83	360,00	237,00	508,00	76,52
Alto preço	696,40	649,00	514,00	993,00	145,90

Ratings

Média	Mediana	Mínimo	Máximo	Desvio-padrão
4,65	4,94	0,00	5,00	1,07

Grupo	Avaliação média
Baixo preço	4,64
Médio preço	4,58
Alto preço	4,83

O grupo com menor avaliação média, mas mesmo assim muito positiva, é o grupo de médio preço, sugerindo que existe um desalinhamento entre expectativas e experiências dos clientes nessa faixa de preço.

Para determinar se essa hipótese é verdadeira, podemos usar o teste ANOVA (Análise de Variância, na sigla em inglês), Esse teste é usado para comparar a média de três ou mais grupos para descobrir se pelo menos um deles tem uma diferença estatisticamente significante dos outros. Se o p-valor for menor que 0,05 (p-value < 0,05), as diferenças são estatisticamente significantes. Os p-valores maiores que 0,05 (p-value > 0,05) sugerem que as diferenças não são estatisticamente significantes e provavelmente são explicadas por aleatoriedade.

Os resultados do teste ANOVA são os seguintes:

- F-statistic: 2.8605275268886468
- p-valor: 0.05797068235613843

Como o p-valor é aproximadamente 0,058 e um pouco maior que o limite, não temos evidências suficientes para rejeitar a hipótese nula (não há diferença entre os valores médios de rating para as faixas de preço) num nível de significância de 5%. Em outras palavras, provavelmente não há diferença significativa entre as avaliações para cada faixa de preço.

Número de avaliações

Média	Mediana	Mínimo	Máximo	Desvio-padrão
78,35	43	0	507	82,31

O número de avaliações, embora interessante, não trouxe grandes insights à compreensão do mercado. Sabemos que o número de acomodações novas (que surgiram nos últimos 3 meses) é de aproximadamente 5%. Levando em consideração somente as acomodações que não são consideradas novas,

Ratings, sem acomodações novas

Levando em consideração somente as acomodações que não são consideradas novas.

Média	Mediana	Mínimo	Máximo	Desvio-padrão
4,90	4,94	4,33	5,00	0,12

Reparem que a média agora ronda muito mais próximo do limite superior.

Grupo	Avaliação média
Baixo preço	4,78
Médio preço	4,93
Alto preço	4,95

A avaliação média por faixa de preço é bastante diferente sem as acomodações novas, bem mais próximas do limite superior.

Número de avaliações, sem acomodações novas

Levando em consideração somente as acomodações que não são consideradas novas.

Média	Mediana	Mínimo	Máximo	Desvio-padrão
82,46	49	3	507	82,41

Observamos que houve mudanças com a retirada das acomodações consideradas novas

Tipo de acomodação

Os tipos de acomodação e sua descrição são os seguintes:

Tipo de acomodação	Proporção	Descrição
Casa	17,7%	A acomodação inteira estará disponível para você. Isso inclui uma entrada privativa e espaços próprios, como um quarto, banheiro e cozinha.
Apartamento	15,9%	--
Quarto	13,8%	Um quarto privativo dentro de uma casa ou apartamento compartilhado. Você terá seu próprio quarto, mas pode compartilhar outras áreas com outros hóspedes ou o anfitrião.
Cabana	13,0%	Uma pequena casa rústica, frequentemente encontrada em áreas florestais ou montanhosas. Oferece uma experiência simples e próxima à natureza, perfeita para uma escapada tranquila.

Chalé	12,9%	Uma casa rústica, frequentemente localizada em áreas montanhosas ou rurais. Oferece uma experiência aconchegante e muitas vezes inclui lareira, varanda e vistas para a natureza.
Loft	6,9%	Um apartamento de conceito aberto, geralmente com um pé direito alto, grandes janelas e um design moderno. Pode ter um mezanino com a cama ou áreas separadas por divisórias.
Microcasa	6,5%	--
Hotel	3,3%	--
Pousada	2,8%	--
Contêiner	1,9%	--
Suíte	1,7%	--
Lugar	1,5%	--
Condomínio	1,5%	--
Trailer	0,3%	--

Aqui estão os 3 principais tipos de acomodação para cada faixa de preço.

	Tipo de acomodação	Proporção (%)	Preço médio
Baixo preço	Quarto	34,8%	R$ 147,70
	Casa	26,7%	R$ 149,60
	Microcasa	9,9%	R$ 215,62
Médio preço	Apartamento	21,5%	R$ 406, 70
	Casa	18,3%	R$ 399,85
	Chalé	13,0%	R$ 345,31
Alto preço	Cabana	37,3%	R$ 709,10
	Chalé	24,2%	R$ 786,26
	Apartamento	13,0%	R$ 660,24

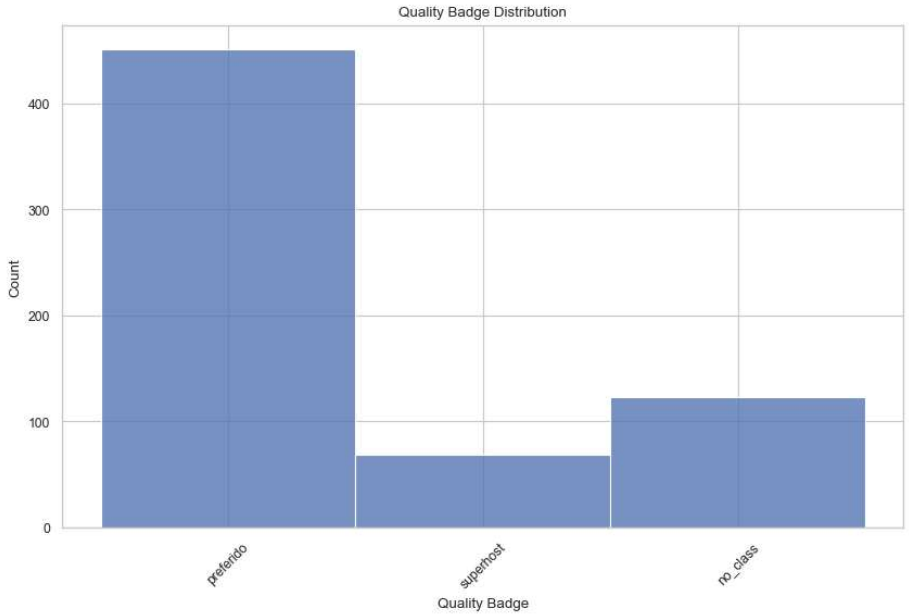
Os quartos são mais comuns na faixa de baixo preço por serem quartos de hotel e pousadas, que estão entre as ofertas de acomodação mais simples e baratas do oferecidos pelo Airbnb. Provavelmente são procurados por clientes mais interessados em aproveitar a

cidade do que a acomodação em si, e que valorizam características como preço baixo, desjejum e distância para o centro turístico e comercial da cidade. Interessante notar também a quantidade de casas inteiras e apartamentos ofertados no Airbnb por um preço relativamente baixo. Na faixa de baixo preço, as casas rivalizam em preço e quantidade com quartos de hotéis e pousadas.

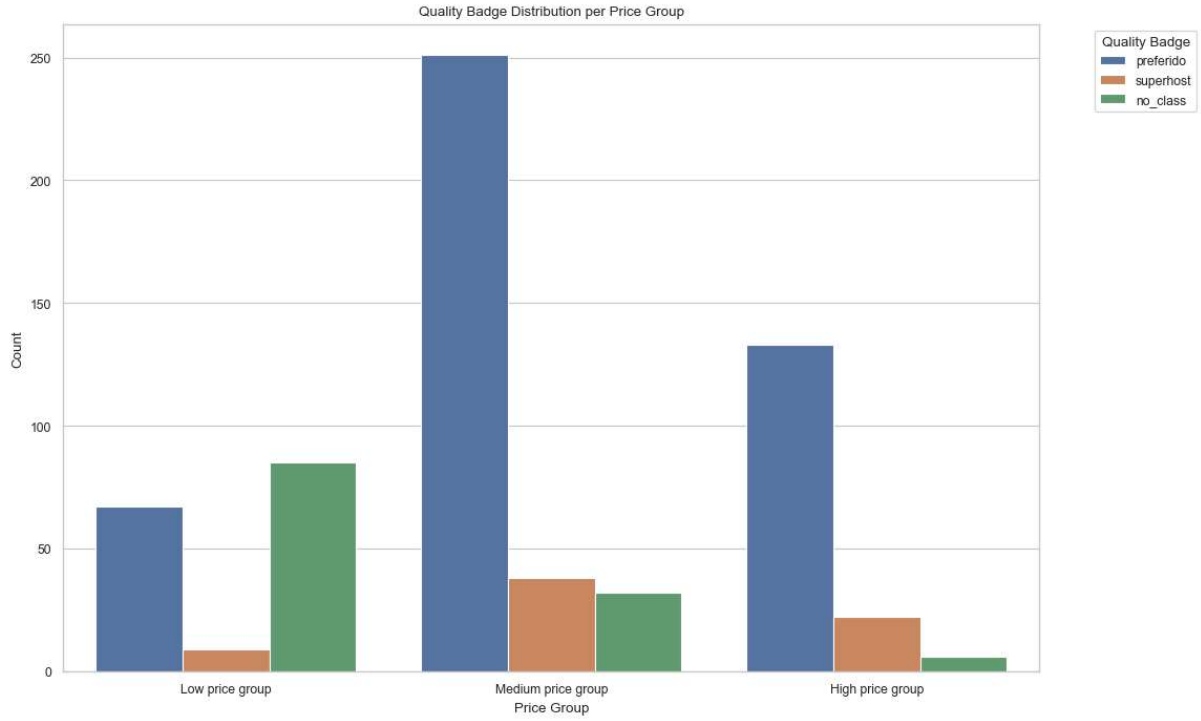
Selos de Qualidade

O elemento coletado das páginas destacavam três tipos de situação como selo de qualidade:

1. **Superhost**: segundo o Airbnb, são anfitriões que fazem o possível e o impossível para oferecer uma excelente hospitalidade. Eles têm mais visibilidade, potencial de ganho e acesso a recompensas exclusivas.
2. **Preferido dos hóspedes**: são acomodações que fazem mais sucesso entre os hóspedes no Airbnb. Possuem excelentes avaliações, com média acima de 4,9 estrelas, e recebem notas altas dos hóspedes em categorias como check-in, limpeza, exatidão do anúncio, comunicação com o anfitrião, localização e custo-benefício.
3. **Sem classificação**: não possuíam selos de qualidade como Superhost ou Preferido dos hóspedes



O problema é que os selos são diferentes e muitas vezes se sobrepõem um ao outro, assim um “Superhost” pode ou não ter acomodações “Preferido dos hóspedes”, por exemplo. Como o selo “Preferido dos hóspedes” é mais difícil de conquistar e manter, ele possui mais valor.



Ao analisarmos a distribuição de selo de qualidade por faixa de preço percebemos que na realidade o termo “Preferido dos Hóspedes” é bastante comum, principalmente no grupo médio, e que é muito mais valoroso a qualidade “Superhost”. Isso não é de se surpreender, já que um dos critérios para receber o título de “Preferido dos Hóspedes” é ter avaliação acima de 4,9 e essa é a avaliação média para os grupos médio e alto.

O título “no_class” se refere ao grupo de anfitriões que não são classificados ou não possuem selos de qualidade. Nesse grupo de anfitriões, como é de se esperar, é mais comum encontrar novas acomodações. Cerca de 18% são de novos anfitriões, em contraste aos 5% do grupo geral. Podemos inferir que a maioria de novas acomodações se encontra na faixa mais acessível.

Amenities

Amenities são as características das acomodações que vão além do tipo. Nela são descritos detalhes como “Vista para as montanhas”, “Estacionamento”, ou “Varanda”. Como não parece existir um modelo ou padrão, no processo de coleta foram encontrados mais de 140 amenities diferentes. Para facilitar a análise dos dados, características semelhantes foram agregadas sob um mesmo título, assim “TV HD”, “TV 48 polegadas”, etc foram reunidas sob a característica “TV”. Dessa maneira, as amenities mais comuns são:

- Estacionamento: 64,07%
- Wifi: 63,00%
- Alarme Monóxido de Carbono: 54,28%

- Detector de fumaça: 52,10%
- Ar-condicionado: 49,00%
- Cozinha: 48,37%
- Televisão (comum, não HDTV): 39,03%
- Vista: 39,81%
- Banheira: 17,73%
- Pátio: 15,24%
- Secador de cabelo: 14,93%

Vale notar algumas características citadas nominalmente pelo nosso cliente:

- Desjejum: 2,33%
- Lareira: 2,49%
- Geladeira ou Frigobar: 4,66%

3. Análise bivariada

Amenities vs Faixa de preço

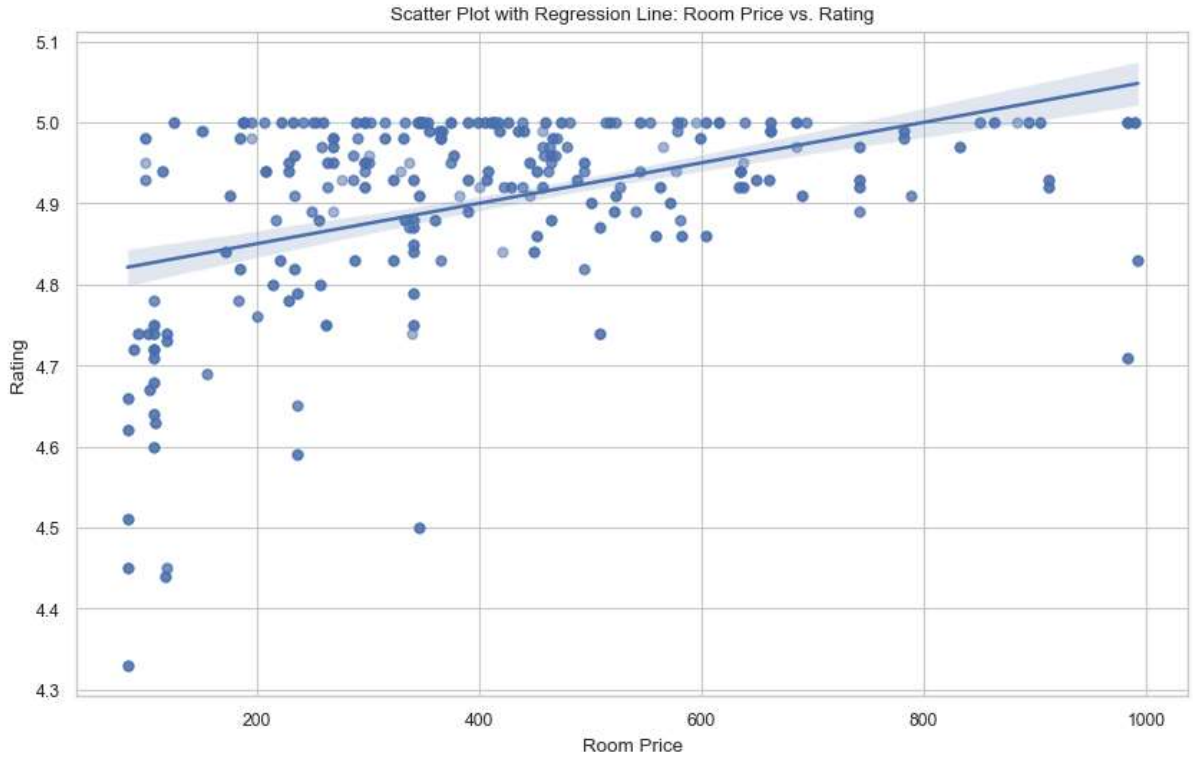
Começando pela proporção de amenities por faixa de preço.

	Acomodação	Proporção
Baixo preço	WiFi	65,84%
	Detector de fumaça	65,21%
	Alarme Monóxido de Carbono	64,60%
	Estacionamento	62,11%
	TV	47,20%

	Acomodação	Proporção
Médio preço	Estacionamento	64,48%
	WiFi	60,75%
	Cozinha	52,71%
	Ar-condicionado	50,78%
	Alarme Monóxido de Carbono	50,15%

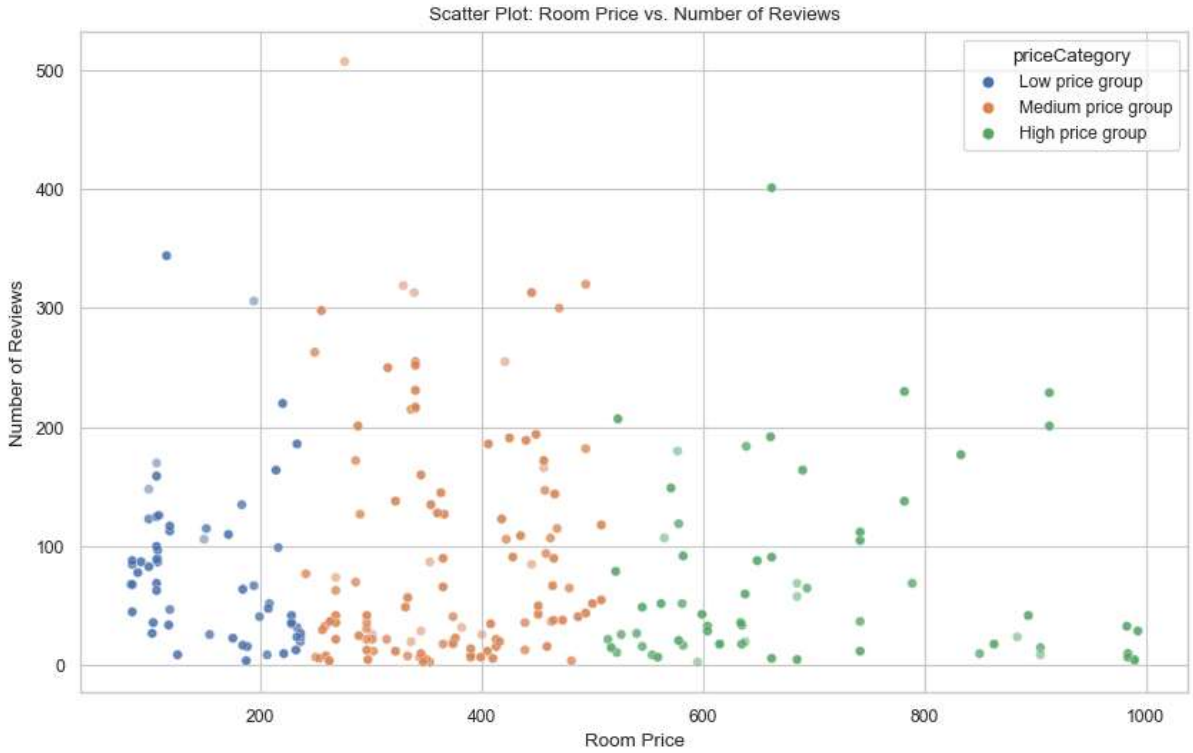
	Acomodação	Proporção
Alto preço	Estacionamento	65,22%
	WiFi	64,60%
	Cozinha	58,39%
	Ar-condicionado	54,04%
	Alarme Monóxido de Carbono	52,17%

Preço vs Ratings (sem acomodações novas)



Existe uma correlação positiva entre rating e o preço praticado.

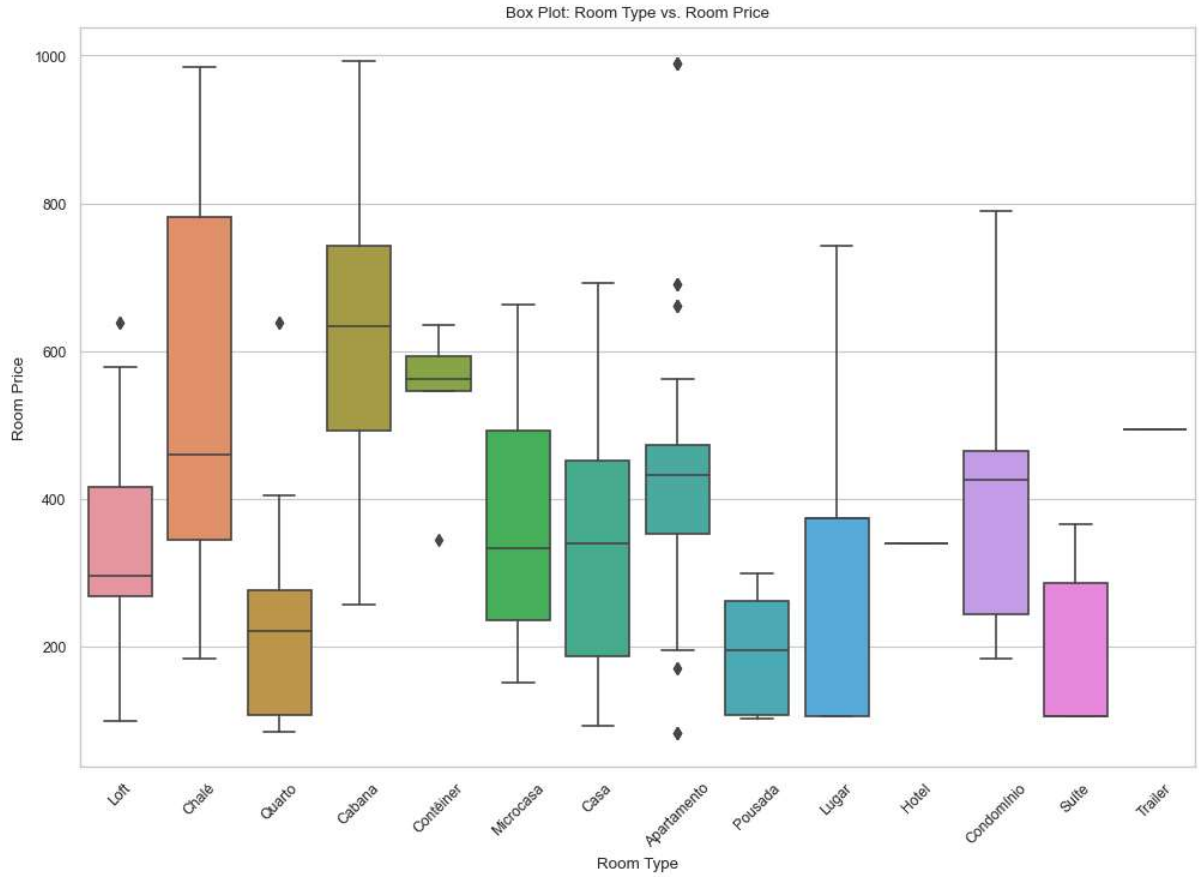
Preço vs Número de Avaliações (sem acomodações novas)



Não existe correlação entre preço e número de avaliações.

A reflexão aqui é que incentivar uma quantidade maior de avaliações não necessariamente irá se refletir num valor maior por diária. Na realidade, provavelmente não terá nenhum impacto.

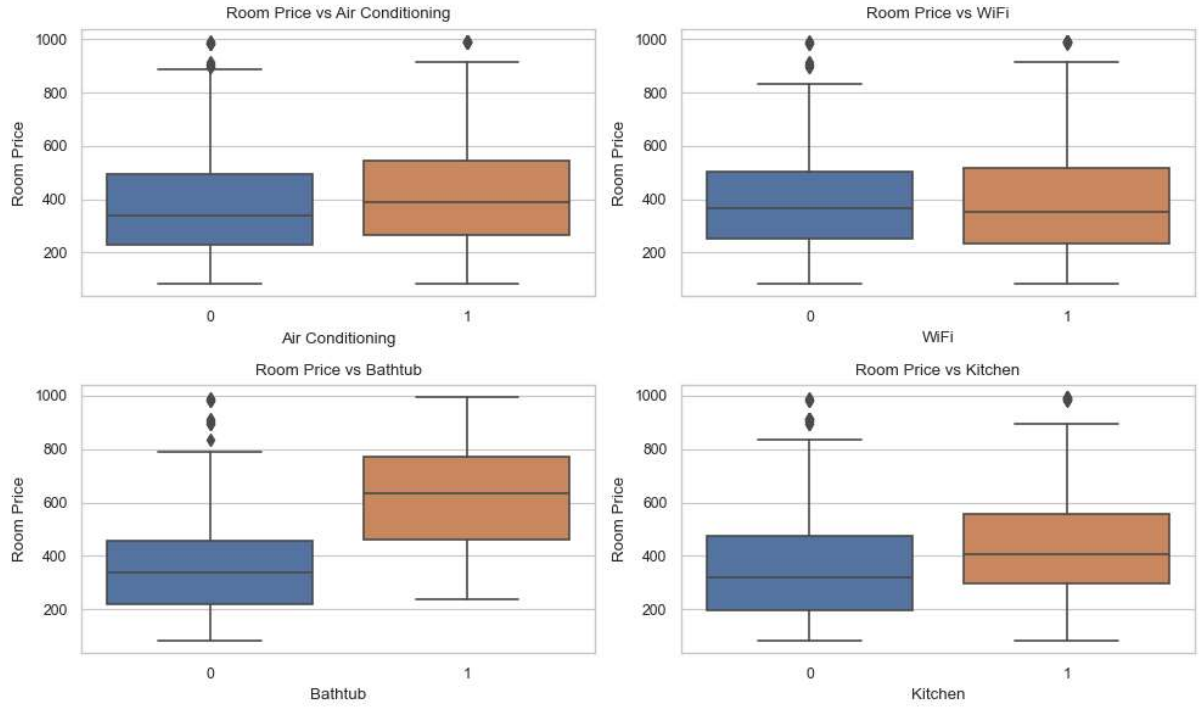
Preço vs Tipo de Acomodação



As acomodações mais rentáveis são Chalé e Cabana, embora possuam uma distribuição bastante grande de preços. As menos rentáveis são Quarto e Pousada, possivelmente refletindo os quartos em hotéis e pousadas mais simples, com preços mais acessíveis.

Os Apartamentos são os que possuem o maior range de preços, com mais outliers, contendo os preços mais altos e mais baixos.

Preço vs Amenities

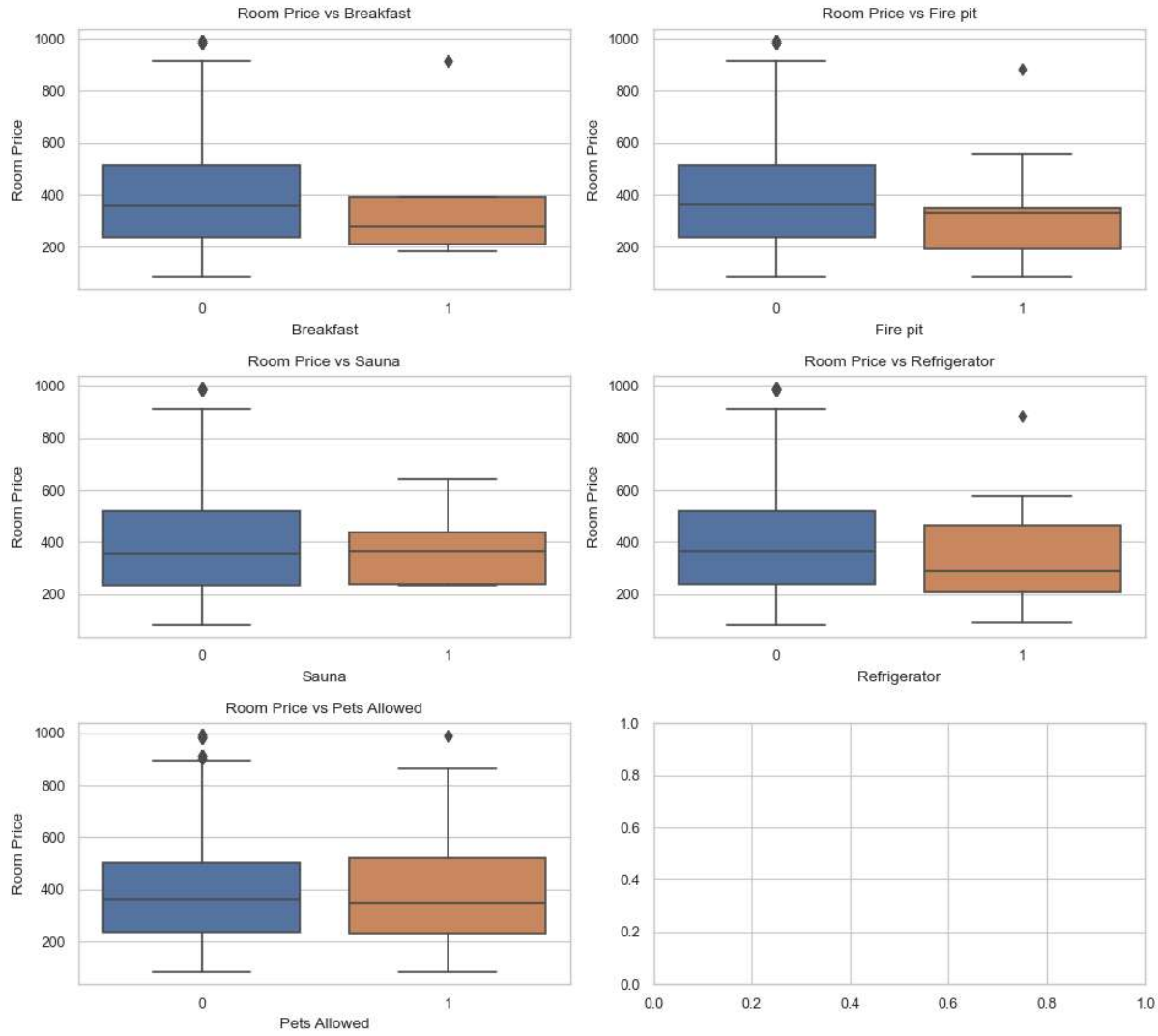


Ar-condicionado: não há diferença de preço significativa.

Banheira: acomodações com banheira são as mais caras.

WiFi: não há diferença de preço significativa.

Cozinha: diferença pequena, com valores levemente mais altos nas acomodações com cozinha.



Desjejum: as acomodações mais acessíveis oferecem desjejum.

Lareira: as acomodações mais acessíveis oferecem desjejum.

Sauna: pequena diferença, mas as acomodações mais acessíveis oferecem desjejum.

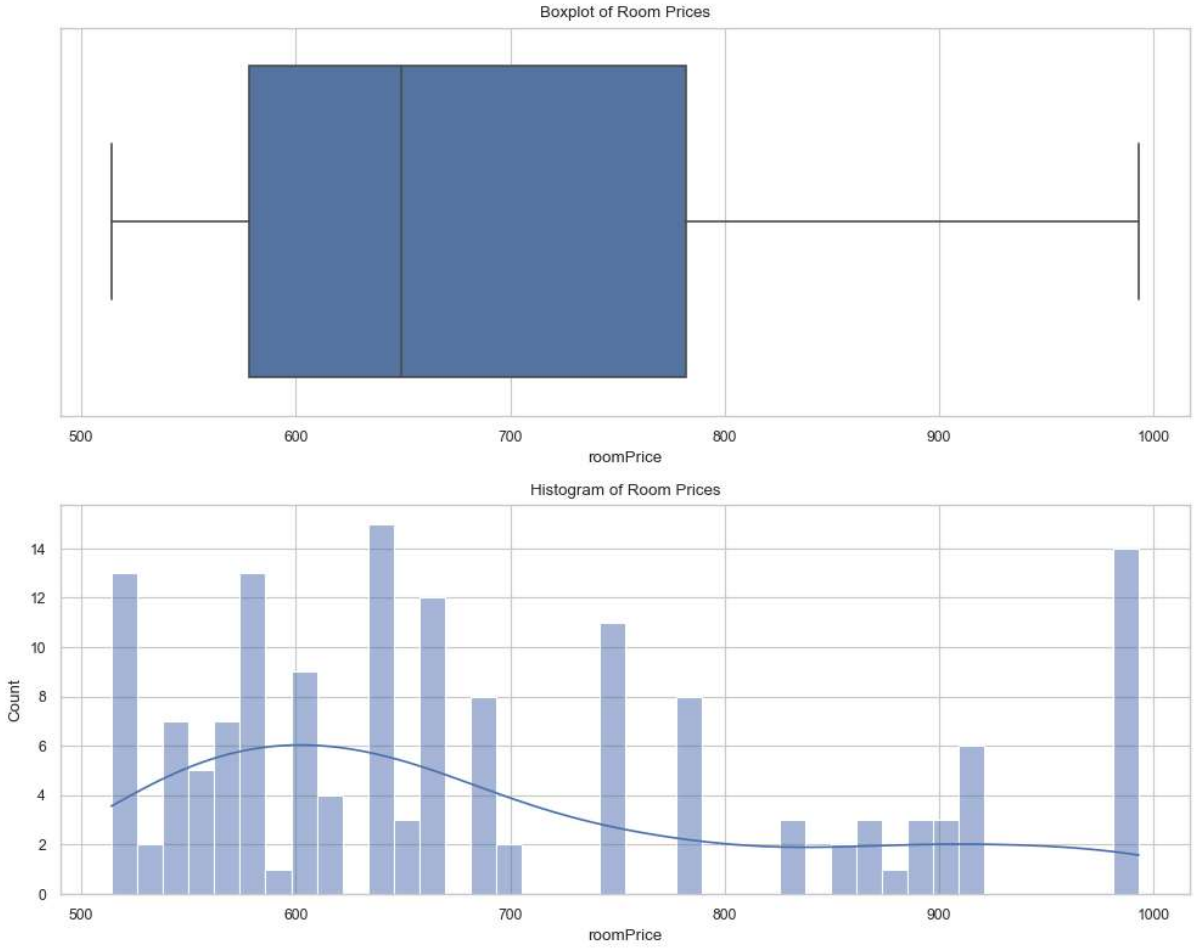
Geladeira/Frigobar: pequena diferença, mas as acomodações mais acessíveis oferecem desjejum.

Pets: não há diferença no preço.

4. Análise da concorrência direta

A pedido do cliente, vamos descrever com mais detalhes um subgrupo de acomodações que ele identifica como sendo seus concorrentes diretos. Esse grupo possui como características o alto valor da diária e ser “Preferido dos Hóspedes” ou “Superhost”.

Preço (R$)

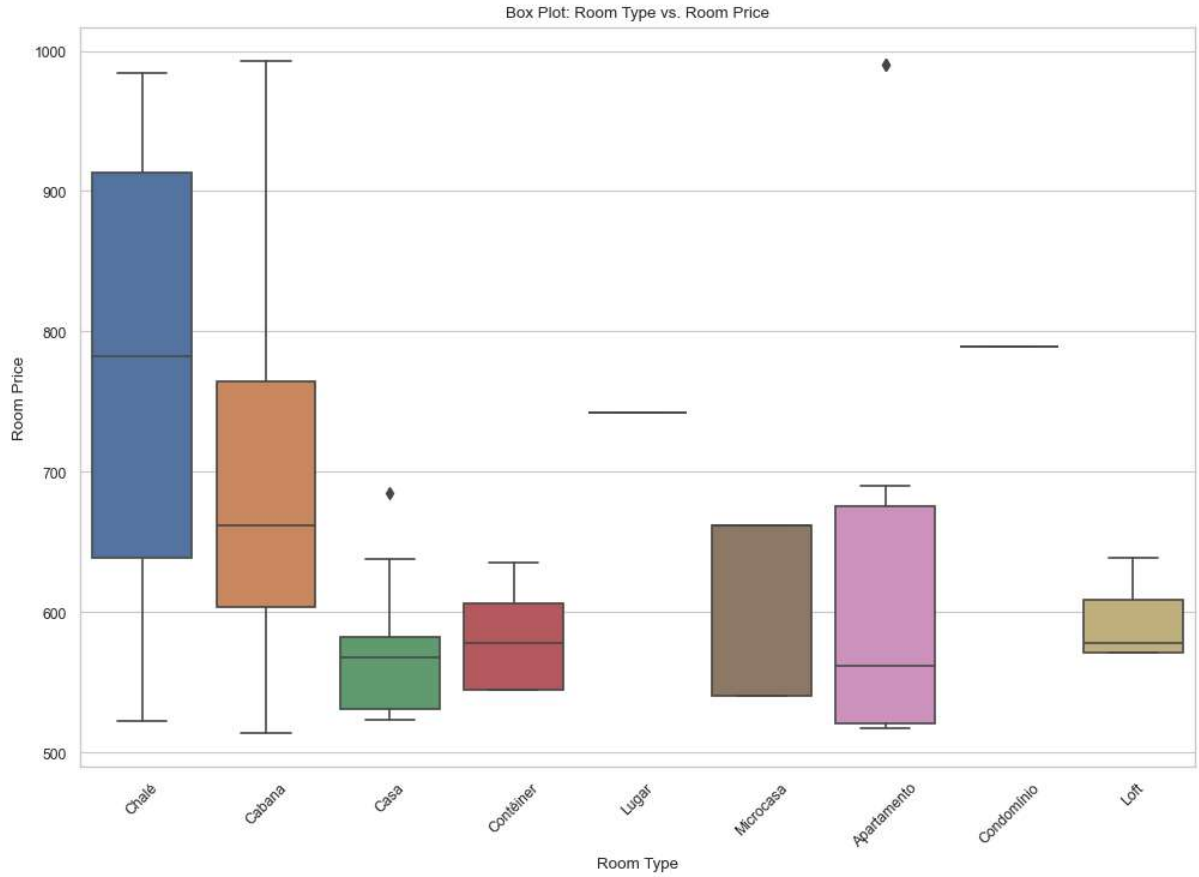


Uma distribuição ainda assimétrica para a direita, com os valores mais comuns ao redor dos R$600,00. Vale a pena citar que não há acomodações novas neste grupo.

Descrição de valores numéricos

	Média	Mediana	Mínimo	Máximo	DP
Preço	693,42	649,00	514,00	993,00	144,72
Ratings	4,95	4,97	4,71	5,00	0,05
Avaliações	74,5	37	3	401	80,67

Tipo de acomodação vs Preço

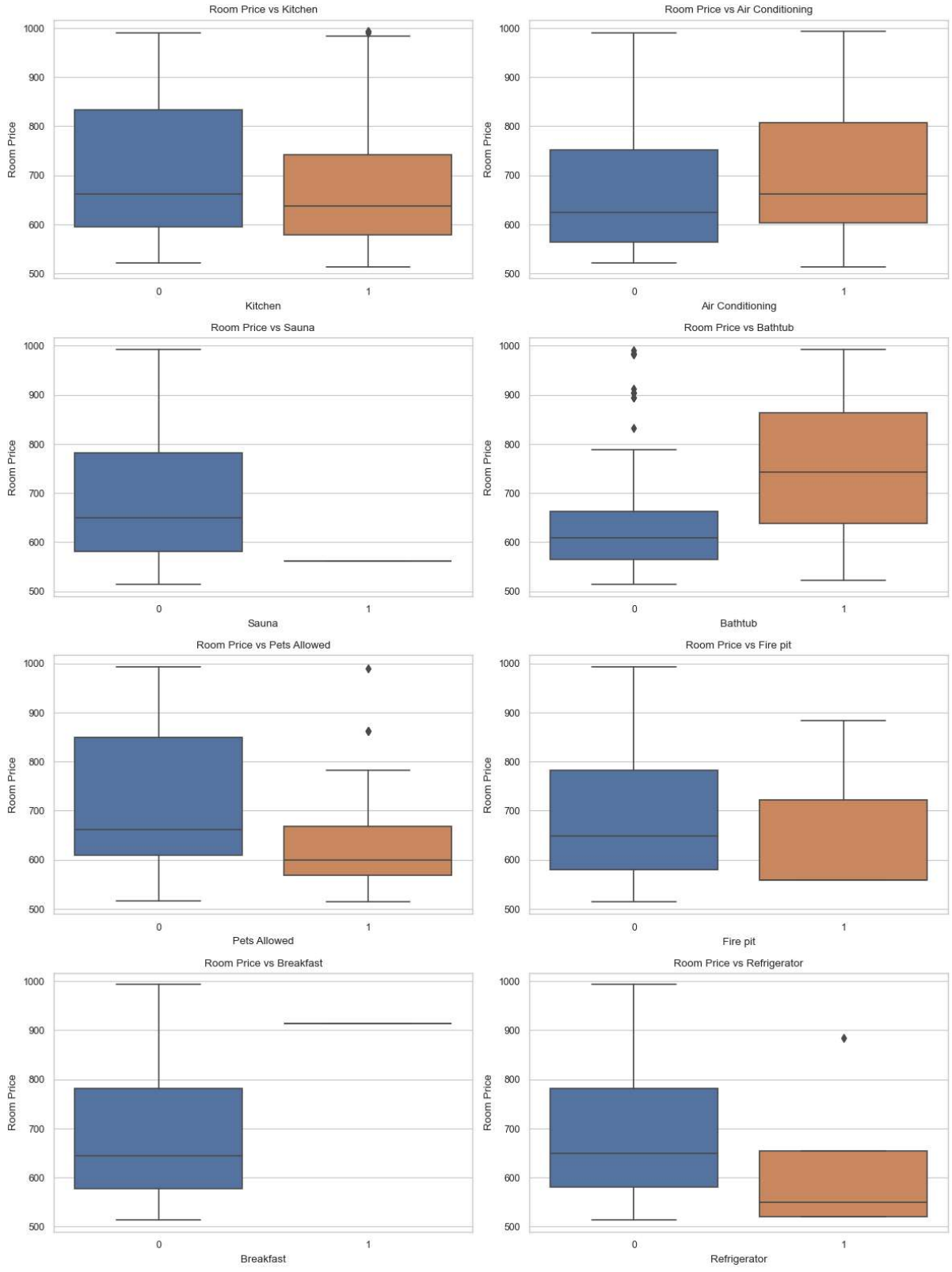


Neste grupo, Chalés e Cabanas continuam sendo os tipos mais comuns, mas também apresentam a maior variância de preço. O tipo com maior mediana é o Chalé.

Amenities

Amenities	Porcentagem
Estacionamento	15,70%
WiFi	15,55%
Cozinha	13,99%
Ar Condicionado	12,91%
Alarme Monóxido de Carbono	12,44%
Alarme de Fumaça	11,04%
Vista (qualquer tipo)	10,73%
Banheira	10,42%
Pets	7,46%

Preço vs Amenities



Cozinha: não há diferença entre a presença de cozinha e preço.

Ar Condicionado: pequena diferença positiva para as acomodações que possuem ar condicionado.

Sauna: acomodações nesse grupo não oferecem sauna.

Banheira: existe uma grande diferença positiva de preço para as acomodações que oferecem banheira.

Pets: as acomodações mais caras não oferecem essa característica.

Lareira: as acomodações mais caras não oferecem essa característica.

Desjejum: acomodações nesse grupo não oferecem desjejum.

Geladeira/Frigobar: as acomodações mais caras não oferecem essa característica.

5. Conclusões e insights

Analisar um mercado tão complexo e dinâmico como o de aluguéis pelo Airbnb em Campos do Jordão é uma tarefa desafiadora. As informações coletadas não estão completas e representam uma amostra da população real de acomodações da cidade. Além disso, elas certamente não representam todos os fatores que podem influenciar nosso objetivo final, que é prever o preço das diárias. Apesar dessas limitações, foi possível chegar a conclusões valiosas, obter insights e delinear os novos passos.

Por exemplo, o crescimento de 5% das ofertas de acomodações é bastante importante. Considerando uma população com cerca de mil ofertas, 50 novas acomodações foram listadas no site nos últimos meses, principalmente na faixa mais acessível. Provavelmente a maioria dessas novas acomodações são quartos de hotéis e pousadas, e não propriedades de pequenos anfitriões.

O selo de qualidade “Preferido dos Hóspedes”, apesar de ser bastante criterioso, é bastante comum. Isso pode refletir um cuidado especial da maioria dos anfitriões, mas também pode significar que ele possui pouco valor e seus critérios de acesso deveriam ser revistos.

Existe uma correlação positiva entre ratings e preços, mas numa faixa muito estreita. A média de avaliação das acomodações é bastante alta e o desvio-padrão é muito pequeno. Provavelmente essas pequenas variações de rating não influenciam os preços.

As amenidades que mais agregam valor são banheira e ar-condicionado.

O preço solicitado pela principal acomodação do cliente está abaixo da mediana do grupo de principais concorrentes. Ele também não oferece banheira e ar-condicionado nesta acomodação.

Há muitas melhorias a serem feitas nesta análise. Vale destacar uma agregação menor das amenidades, o que poderia diferenciar melhor os subgrupos (como os tipos de banheira), uma análise estatística mais avançada, com mais testes estatísticos, e focar a coleta de dados nos concorrentes diretos das acomodações do cliente.

Para os próximos projetos podemos avaliar a coleta e análise de sentimento das avaliações de clientes, recorrência da análise em períodos diferentes do ano e atualização de dados, entre outras sugestões do cliente e seus associados.