**Análise de Dados do Airbnb | Rio de Janeiro**

Vista aérea de água e montanha ao fundo

O conteúdo gerado por IA pode estar incorreto.

Este artigo apresenta um resumo dos principais insights obtidos na minha análise das acomodações do Airbnb do Rio de Janeiro. Para conferir o estudo completo com todos os detalhes, acesse meu repositório no GitHub abaixo:

**1. O que é o Airbnb?**

Fundado em 2008, o Airbnb revolucionou o setor de hospedagem ao conectar viajantes a anfitriões dispostos a alugar suas propriedades. Com uma plataforma prática e inovadora, a empresa se tornou a maior do ramo, mesmo sem possuir um único hotel. Presente em quase todos os países, já recebeu mais de 1,4 bilhão de hóspedes. Apesar do sucesso, a pandemia impactou fortemente o negócio, levando à demissão de 25% de seus funcionários.

**2. Por que o Rio de Janeiro?**

O Rio de Janeiro é, indiscutivelmente, um dos principais destinos turísticos do Brasil e da América Latina. Com suas praias mundialmente famosas, como Copacabana e Ipanema, o Cristo Redentor entre as Sete Maravilhas do Mundo Moderno, e um calendário cultural repleto de eventos como o Carnaval e o Réveillon, a cidade atrai milhões de visitantes todos os anos.

Esse alto fluxo turístico se reflete diretamente na oferta de hospedagens alternativas, como o Airbnb. De acordo com dados do Ministério do Turismo e de plataformas de hospedagem, o Rio é a cidade brasileira com maior número de imóveis cadastrados no Airbnb — superando capitais maiores em população, como São Paulo.

Analisar os dados do Airbnb no Rio permite entender padrões de preços, comportamento dos anfitriões, distribuição geográfica das propriedades e tendências do mercado de hospedagem compartilhada em um dos cenários mais movimentados do Brasil. Além disso, oferece insights valiosos tanto para viajantes quanto para investidores e gestores públicos.

**3. Obtenção e Conhecimento Inicial dos Dados**

O Airbnb disponibiliza dados abertos de diversas cidades pelo mundo através do portal [Inside Airbnb](http://www.insideairbnb.com/get-the-data.html). Para esta análise exploratória, utilizei a versão resumida da base referente ao Rio de Janeiro, ideal para visualizações:

Mapa

O conteúdo gerado por IA pode estar incorreto.

**Arquivo utilizado:** listings.csv – resumo com métricas e informações principais dos anúncios.

**4. Conhecimento e Pré-Processamento dos Dados**

**Conhecendo os dados**

* id - número de id do imóvel
* name - nome da propriedade anunciada
* host\_id - número de id do proprietário da hospedagem
* host\_name - Nome do proprietária
* neighbourhood\_group - esta coluna não contém nenhum valor válido
* neighbourhood - nome do bairro
* latitude - coordenada da latitude da propriedade
* longitude - coordenada da longitude da propriedade
* room\_type - informa o tipo de quarto que é oferecido
* price - preço para alugar o imóvel
* minimum\_nights - quantidade mínima de noites para reservar
* number\_of\_reviews - número de reviews que a propriedade possui
* last\_review - data do último review
* reviews\_per\_month - quantidade de reviews por mês
* calculated\_host\_listings\_count - quantidade de imóveis do mesmo anfitrião
* availability\_365 - número de dias de disponibilidade dentro de 365 dias

**Primeiras Entradas**

Antes de iniciar qualquer análise, é fundamental observar as primeiras entradas do conjunto de dados. Utilizando o método .head() do Pandas, conseguimos visualizar as cinco primeiras linhas do DataFrame e ter uma noção inicial da estrutura das informações.

Tela de computador

O conteúdo gerado por IA pode estar incorreto.

Já é possível notar que temos uma coluna com vários valores ausentes (NaN), vamos tratar desse assunto um pouco mais adiante.

**Linhas e colunas do banco de dados**

Além de visualizar as primeiras entradas com o método .head(), utilizamos o método .shape para identificar a dimensão do conjunto de dados, ou seja, a quantidade de variáveis (colunas) e registros (linhas).

O dataset resumido utilizado nesta análise contém 18 colunas e 39.499 registros. A partir dessas variáveis, buscaremos extrair os principais insights e padrões relevantes. Essa etapa inicial é essencial para compreender a estrutura da base e direcionar os próximos passos da análise exploratória.

Dimensões do Dataset

**Dimensões do dataframe**

O método info() é util para a obtenção de uma rápida descrição dos dados, em especial o número total de linhas, o tipo de cada atributo e o número de valores não nulos.

Texto

O conteúdo gerado por IA pode estar incorreto.

Dos 16 atributos presentes no dataset, 11 são do tipo numérico (inteiros ou decimais), enquanto os demais são do tipo object.

**Atributos do tipo texto**

Entre as colunas do tipo texto, destacam-se neighbourhood e room\_type, cujos valores se repetem com frequência. Isso indica que provavelmente se tratam de variáveis categóricas. Para confirmar essa hipótese e verificar quais categorias estão presentes, utilizamos o método .value\_counts().

Texto

O conteúdo gerado por IA pode estar incorreto.

**Atributos do numéricos**

Agora que já conhecemos as dimensões do nosso dataset, é hora de explorar algumas estatísticas descritivas. Vamos focar apenas nas variáveis numéricas que serão analisadas ao longo do artigo: price, minimum\_nights, number\_of\_reviews e reviews\_per\_month. Esses atributos fornecem informações importantes sobre preços, tempo mínimo de estadia e o nível de engajamento dos hóspedes com os anúncios.

Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

**Preço (price)**

* **Média:** O preço médio dos imóveis é R$ 1.227, mas com um desvio padrão alto de R$ 3.766, o que indica que existem grandes variações nos preços e indicando a presença de **outliers**
* **Mínimo:** O preço mínimo é R$ 33, o que é bastante baixo, indicando a existência de algumas ofertas muito baratas.
* **Máximo:** O preço máximo é R$ 500.000, sugerindo que existem imóveis de luxo ou muito exclusivos na base de dados.

**Percentis:**

* 25% dos imóveis têm preço até R$ 400.
* 50% (mediana) dos imóveis têm preço até R$ 728.
* 75% dos imóveis têm preço até R$ 1.200, o que indica que a maioria dos imóveis fica em uma faixa de preço de até R$ 1.200.

**Noite Mínima (minimum\_nights)**

* **Mínimo:** O valor mínimo de 1 noite indica que existem imóveis disponíveis para aluguel de curto período.
* **Máximo:** O valor máximo de 1000 noites provavelmente indica imóveis com requisitos de reserva muito longos ou listagens com erros. É evidente que esse valor foge completamente do padrão dos demais dados, indicando a presença de um **outlier**. Esses valores extremos podem distorcer a análise e, por isso, devem ser tratados com cautela.

**Qual o tipo de distribuição das variáveis?**

 Uma forma rápida de visualizar isso é criando um histograma para cada variável numérica. Podemos fazer isso individualmente ou usar o método .hist() para todo o conjunto de dados.

Interface gráfica do usuário, Calendário

O conteúdo gerado por IA pode estar incorreto.

**Verificando valores nulos**

o utilizar o método isnull(), conseguimos identificar os valores ausentes. Para contar o número exato de valores faltantes por variável, aplicamos o método .sum(). Como o objetivo é calcular a porcentagem de dados ausentes em cada coluna, dividimos o número de valores nulos pelo total de linhas do dataset, usando df.shape[0], onde df é o seu DataFrame. A seguir, apresento o resultado, com as variáveis ordenadas em ordem decrescente.

Texto

O conteúdo gerado por IA pode estar incorreto.

**Altos percentuais de valores nulos:**

* last\_reviewe reviews\_per\_month têm 26,38% de valores nulos, o que é um percentual considerável. Isso indica que muitas propriedades podem não ter recebido avaliações recentes ou mensais, o que pode ser esperado para novos imóveis ou para aqueles com pouca visibilidade.

**Baixo percentual de valores nulos:**

* pricetem apenas 7,23% de valores nulos, o que é relativamente baixo, mas ainda assim deve ser tratado, já que o preço é uma variável importante para análise de dados.
* host\_nametem um percentual de 0,005%, indicando que a grande maioria dos registros tem informações sobre o nome do anfitrião.

**Tratando outliers**

**Definição de Funções para Análise de Outliers**

Antes de analisar as variáveis numéricas mais relevantes, é importante garantir que os dados estejam limpos e representem bem a realidade. Um dos principais cuidados nesse processo é lidar com os *outliers* — valores que fogem completamente do padrão da maioria dos dados.

Outliers são pontos extremos que podem surgir por erro de digitação, exceções raras ou características muito específicas de um imóvel. Por exemplo, um anúncio com 1.000 noites mínimas de reserva ou um preço absurdamente alto pode distorcer médias, gráficos e qualquer análise estatística.

Uma forma prática de identificá-los é utilizando o intervalo interquartil (IQR), um método simples e eficaz. O processo envolve calcular:

* primeiro quartil (Q1), que representa os 25% menores valores, e
* terceiro quartil (Q3), que representa os 75%.

A diferença entre eles nos dá o IQR, que serve como base para definir os limites aceitáveis. A partir disso, qualquer valor abaixo de **Q1–1.5 × IQR** ou acima de **Q3 + 1.5 × IQR** é considerado um outlier e pode ser removido do dataset.

Esse tipo de limpeza ajuda a evitar distorções em análises estatísticas, garantindo que nossos resultados sejam mais consistentes e realistas.

def limites(coluna):  
 q1 = coluna.quantile(0.25) # Primeiro quartil  
 q3 = coluna.quantile(0.75) # Terceiro quartil  
 amplitude = q3 - q1 # Amplitude interquartil  
 return q1 - 1.5 \* amplitude, q3 + 1.5 \* amplitude # Limites inferior e superior  
  
def excluir\_outliers(df, nome\_coluna):  
 qtde\_linhas = df.shape[0] # Pega a quantidade de linhas atuais  
 lim\_inf, lim\_sup = limites(df[nome\_coluna]) # Calcula os limites inferior e superior para a coluna  
 df = df.loc[(df[nome\_coluna] >= lim\_inf) & (df[nome\_coluna] <= lim\_sup), :] # Filtra os dados  
 linhas\_removidas = qtde\_linhas - df.shape[0] # Calcula o número de linhas removidas  
 return df, linhas\_removidas # Retorna o DataFrame filtrado e o número de linhas removidas

**Excluindo atributos irrelevantes**

Durante a análise exploratória, percebi que as variáveis price e minimum\_nights continham valores extremamente altos, destoando do padrão geral do conjunto. Para investigar melhor, utilizei duas abordagens rápidas:

* resumo estatístico com describe()
* visualização por meio de boxplots.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Os boxplots são ótimas ferramentas para identificar outliers, que aparecem como pontos fora da caixa. Esses valores extremos não só distorcem a percepção visual — comprimindo os quartis — como também influenciam medidas como a média, comprometendo a interpretação dos dados.

Diante disso, optei por tratar esses outliers para garantir análises mais representativas da realidade. Para evitar conflitos com execuções anteriores, primeiro criei uma cópia do DataFrame original:

# Cópia dos dados originais  
df\_original = base\_airbnb.copy()  
  
# Limpeza de outliers com suas funções  
df\_limpo = df\_original.copy()  
df\_limpo, \_ = excluir\_outliers(df\_limpo, 'price')  
df\_limpo, \_ = excluir\_outliers(df\_limpo, 'minimum\_nights')

Após a limpeza, replotei os histogramas para observar as distribuições sem a interferência dos outliers. As mudanças visuais foram significativas, principalmente nas colunas price e minimum\_nights.

Gráfico, Histograma

O conteúdo gerado por IA pode estar incorreto.

Para complementar essa análise, comparei a média das variáveis antes e depois da remoção dos outliers, o que permitiu observar com mais clareza o impacto estatístico desses valores extremos. Após a limpeza do DataFrame, foi possível notar que:

* As médias das variáveis price e minimum\_nights se aproximaram significativamente da mediana (50%), indicando uma distribuição mais equilibrada;
* Os histogramas passaram a representar melhor a distribuição real dos dados, facilitando a visualização e interpretação.

**5. Análise Exploratória de Dados**

Nesta etapa vamos realizar uma análise exploratória para conhecermos os dados. Isto é fundamental para entender o comportamento dos dados, entender quais atributos são importantes e quais podem ser descartados. É nessa fase também que vamos realizar a limpeza nos dados caso seja necessário.

**Correlação entre variáveis**

Correlação indica a existência de uma relação entre duas variáveis. No contexto atual, estamos analisando a relação ou similaridade entre essas variáveis.

Essa relação pode ser quantificada por meio do coeficiente de correlação, que determina a intensidade dessa conexão. Para identificar as correlações entre as variáveis de interesse, vou:

* Construir uma matriz de correlação
* Gerar um heatmap a partir dessa matriz utilizando a biblioteca seaborn
* Identificação das relações entre diferentes atributos.

Gráfico, Gráfico de mapa de árvore

O conteúdo gerado por IA pode estar incorreto.

Através deste Heatmap, observamos apenas uma correlação fraca entre number\_of\_reviewse reviews\_per\_month.

**Qual o tipo de imóvel mais comum?**

Por meio do método value\_counts() , verificou-se que:

* Aproximadamente 80% dos aluguéis do Airbnb, no Rio de Janeiro, são de casas ou apartamentos inteiros;
* 20% das propriedades são quartos privativos;
* 1% dos tipos de aluguéis representam Quartos de hotel e quartos compartilhados, juntos

Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

**Qual a média de preços por tipo de quarto?**

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

**Qual a região com mais imóveis para alugar?**

para descobrir qual a quantidade de ofertas por bairro, no gráfico vou mostrar as 10 primeiras localidades em oportunidades.

Uma imagem contendo Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

**Qual da média de preços por região?**

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

**Mapa de calor dos preços**

Visualização espacial da distribuição dos preços.

Mapa, Gráfico de dispersão

O conteúdo gerado por IA pode estar incorreto.

**Há correlação entre mínimo de noites e preço?**

ara verificar **se há correlação entre o número mínimo de noites (minimum\_nights) e o preço (price)**, você pode calcular a **correlação de Pearson**, que mede o grau de relação linear entre duas variáveis.

O valor de correlacao.loc['minimum\_nights', 'price'] indica a força e direção da correlação:

* **Próximo de 1**: Correlação forte positiva (quanto mais noites, maior o preço).
* **Próximo de -1**: Correlação forte negativa.
* **Próximo de 0**: Sem correlação linear.

Texto

O conteúdo gerado por IA pode estar incorreto.

Visualização com gráfico de dispersão (scatter plot):

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

1. O preço não depende fortemente da quantidade mínima de noites.

* Ou seja, **minimum\_nights e price têm correlação fraca ou nula**, como provavelmente confirmado pelo coeficiente de correlação de Pearson perto de 0.

2. Há grande variabilidade de preços mesmo com a mesma regra de mínimo de noites.

* Isso sugere que outros fatores (como localização, tipo de imóvel, capacidade, avaliação etc.) são mais determinantes no preço.

Uma Análise mais clara, podemos calcular o **preço médio por quantidade mínima de noites**

Interface gráfica do usuário, Aplicativo, Teams

O conteúdo gerado por IA pode estar incorreto.

Ao observarmos a dispersão dos dados, notamos que, para cada valor de minimum\_nights, há uma ampla variação nos preços, formando colunas verticais. Isso por si só sugere **baixa correlação direta** entre as variáveis.

Contudo, ao calcular a **média de preço por quantidade mínima de noites**, observa-se uma **tendência de crescimento do preço médio até 3 noites**. A partir desse ponto, o valor médio tende a se estabilizar em torno de **R$ 900**, com pequenas variações.

* **1 a 3 noites**: Pode refletir imóveis com alta rotatividade (curta temporada), como studios e apartamentos em áreas turísticas, que praticam preços mais acessíveis.
* **Acima de 3 noites**: Normalmente são locações com um perfil mais prolongado, e os preços parecem seguir um padrão mais estável, possivelmente negociado ou ajustado para atratividade.

**6. Principais Insights e Interpretações**

**Comparações interessantes entre bairros e tipos de imóvel**

**Tendências sazonais e comportamentos turísticos**

**Comportamento dos anfitriões profissionais vs. individuais**

**7. Conclusões Finais**