RELATÓRIO TÉCNICO: ANÁLISE DE RECEITA AIRBNB NY

por Gabrielle Rosa Ribeiro e Nayara Ramos de Santana

Título: Análise de Fatores de Receita para Acomodações do Airbnb em Nova York (Foco em hospedagens de Alta Performance)

Nome da base: New York City Airbnb Open Data (Kaggle e Inside Airbnb)

1. Contexto e Objetivo

O mercado de aluguéis por temporada em Nova York é altamente competitivo e regulamentado (especialmente para curta duração < 30 dias). Para o Airbnb e seus anfitriões, entender os fatores que impulsionam o sucesso financeiro – definido por alta ocupação e capacidade de comandar bons preços – é crucial para otimizar estratégias e maximizar a receita.

O objetivo desta análise é identificar as características chave das acomodações e os comportamentos dos anfitriões que estão mais associados a uma maior receita anual estimada, focando especificamente no segmento de alta performance (acomodações com alta ocupação estimada e alta demanda futura indicada pela disponibilidade). O objetivo final é derivar recomendações estratégicas acionáveis para o Airbnb aumentar a receita geral na plataforma em NY.

1.1. Pergunta Central

Quais são os principais fatores (características da acomodação e perfil do anfitrião) que impulsionam a receita das listagens do Airbnb em Nova York, e como esses insights podem ser traduzidos em estratégias eficazes para aumentar a receita da plataforma na cidade?

1.2. Hipóteses que serão investigadas

- Quais localizações (bairros) são mais rentáveis?
- Qual o impacto do tipo/tamanho da acomodação?
- Quais notas de avaliação se correlacionam com o preço?
- Quais comodidades agregam mais valor? A quantidade de comodidades influenciam na escolha dos hóspedes?
- Como a duração mínima da estadia impacta a receita anual? Considerando as regulações de NY.
- Anfitriões Superhost geram mais receita?
- A responsividade/aceitação do anfitrião impacta a performance?
- A experiência (tempo na plataforma) do anfitrião influencia nos resultados?
- A verificação de identidade tem impacto?

2. Fontes de Dados e LGPD

As bases de dados utilizadas foram obtidas de fontes públicas online:

I. **Base Principal**: Dgomonov (Data Publisher). New York City Airbnb Open Data. Kaggle Dataset. Acessado em [14/10/2025], de: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data.

II. **Base Complementar** (Listings Detalhadas): Inside Airbnb. Get the Data - New York City. Acessado em [14/10/2025], de: http://insideairbnb.com/get-the-data/ (Especificamente o arquivo "listings.csv.gz" - detailed listings).

As bases de dados do Kaggle e Inside Airbnb são compiladas a partir de informações publicamente disponíveis na plataforma Airbnb e são amplamente utilizadas pela comunidade acadêmica e de análise de dados para fins de estudo e pesquisa.O uso dessas bases neste projeto tem finalidade estritamente educacional e analítica, visando compreender tendências de mercado no setor de hospedagem.

Os dados utilizados não contêm dados pessoais sensíveis conforme definidos pela Lei Geral de Proteção de Dados (LGPD). Informações que poderiam identificar diretamente indivíduos (como nomes completos de anfitriões e hóspedes, dados de contato exatos) não estão presentes ou foram anonimizadas nas fontes utilizadas (ex: IDs numéricos). O tratamento dos dados se limita à análise agregada de padrões e tendências, sem foco em indivíduos específicos, e ocorre de forma transparente dentro deste projeto educacional. A coleta original dos dados pelo Airbnb e sua disponibilização pública (parcialmente agregada/anonimizada) pelas plataformas Kaggle e Inside Airbnb pressupõem consentimento ou base legal para tal, e nosso uso se enquadra na finalidade legítima de análise de dados abertos para fins educacionais.

3. Processo de Análise de dados

O processo seguiu etapas padrão de análise de dados, utilizando Databricks (PySpark) para ETL e Power BI para modelagem e visualização:

a) **Compreensão das bases:** Fizemos um estudo inicial dos dados antes de partir para a prática usando um plugin da plataforma Siyuan. Encontramos 4 tipos principais de informações:

- de anfitriões
- de localização
- de acomodações listadas
- de review

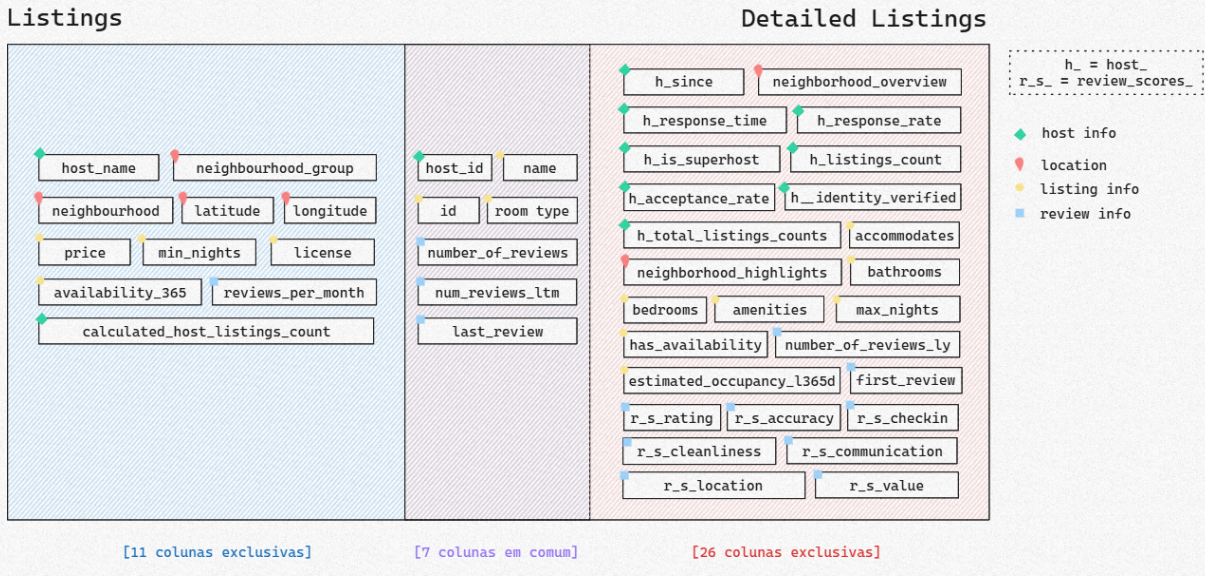

imagem 1 - representação da intersecção de colunas dos 2 datasets separando visualmente os tipos de informações encontradas

Neste cenário, ficou clara para nós a necessidade de uma segregação dos dados em diferentes tabelas, para uma análise mais consistente e fácil de visualizar, além de colunas com informações insuficientes ou desnecessárias para o caráter da nossa investigação (img. 2).

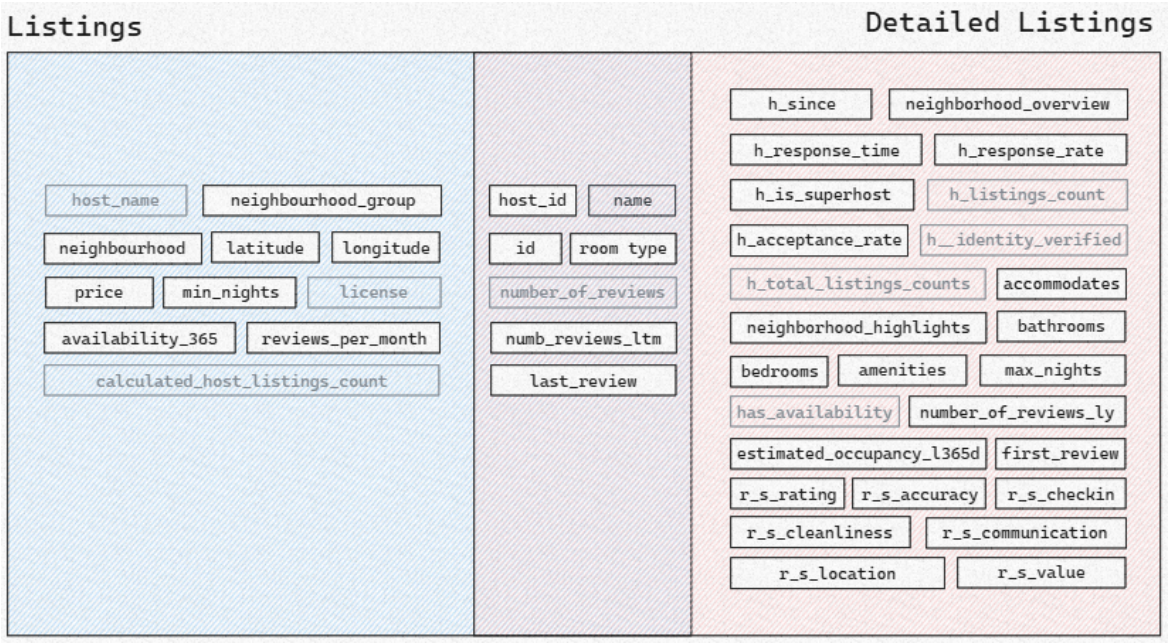

imagem 2 - representação da intersecção de colunas dos 2 datasets destacando apenas colunas selecionadas

Fizemos um projeto inicial de modelagem (img.3) para organizar a separação de tabelas e então poder ir para o databricks. A categoria amenities ganhou uma tabela própria por ser uma lista de comodidades que decidimos converter para colunas booleanas em buscar uma correlação com o faturamento ou taxa de ocupação.

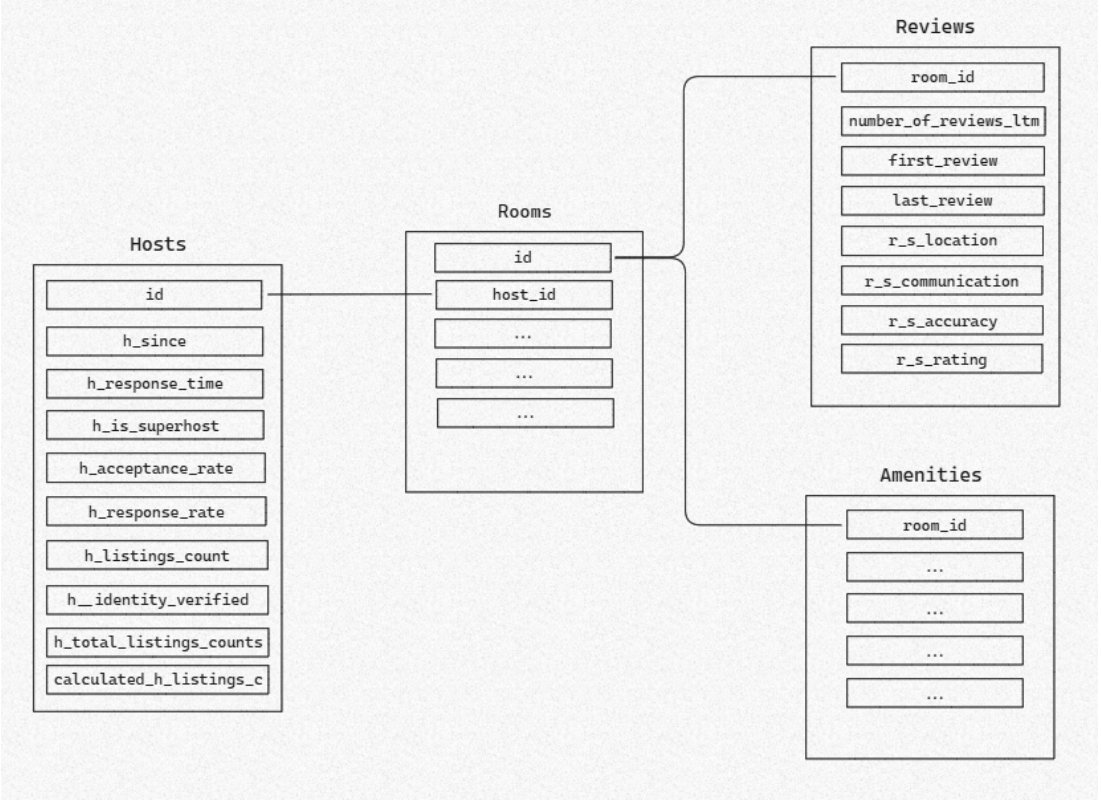

imagem 3 - rascunho inicial da modelagem das tabelas

b) **Aquisição e Consolidação:** Leitura dos dois arquivos CSV, investigação da chave (`id`), correção de formatos inconsistentes (notação científica), remoção de colunas duplicadas e junção (`inner join`) para criar um DataFrame unificado (`df_bnb`).

c) **Limpeza Geral e Seleção (df_final):**

- Substituição global de “N/A” por valores nulos.
- Casting de tipos inicial (colunas que não precisam de tratamento)
- Análise extensiva da distribuição de nulos.

- Segmentação exploratória por atividade de review (`number_of_reviews_ltm`) e disponibilidade de preço (`price`).
- Decisão estratégica para definição do dataset final de análise (`df_final`) como a união de acomodações com preço válido ou com review nos últimos 30 meses, para criar um dataset abrangente focado em registros relevantes.

d) Normalização: Divisão do `df_final` em quatro tabelas lógicas para tratamento individual: `rooms`, `hosts`, `reviews`, `amenities`, sem incluir as colunas não relevantes para a análise.

e) Limpeza Detalhada e Sincronização:

- **`df_rooms`:** Casting de tipos (`bathrooms` com criação de `incomplete_bathroom`, `latitude`, `longitude`), remoção sistemática de outliers de preço via método IQR e remoção de valor mínimo implausível ($3), imputação de `neigborhood_overview` (“Not informed”) para valores ausentes.
- **Sincronização:** Após limpeza do `df_rooms`, os DataFrames `df_hosts`, `df_reviews`, `df_amenities` foram filtrados (via `semi join`) para garantir integridade referencial, mantendo apenas registros correspondentes aos `id`s e `host_id`s presentes no `df_rooms` final.
- **`df_hosts`:** Verificação de duplicatas por `host_id` na criação, casting de tipos (para informações de taxas para `double` 0-1, booleans para `True/False`), imputação de nulos (response time para “Not Informed”, rates para `0`, booleans para `False`).
- **`df_reviews`:** Casting de `reviews_per_month` para `double`, imputação condicional de nulos (scores, rate, dates) para `0` onde `number_of_reviews` é `0`.

- **`df_amenities`:** Extração e Transformação da coluna amenities (originalmente uma string formatada como lista). Uso de funções PySpark (`regexp_replace`, `split`, `explode`) para limpar caracteres indesejados, separar cada comodidade individualmente e criar uma tabela longa (uma linha por id de quarto e comodidade). Uso a biblioteca MLib para treinar um modelo com os dados separados e usá-lo para categorizar as comodidades mais frequentes/relevantes. Foi feito o pivoteamento (`groupBy`, `pivot`) para criar colunas individuais para cada comodidade principal, preenchidas com 1 (presente) ou 0 (ausente). Criação de uma coluna adicional (`amenities_count`) contando o número total de comodidades por quarto.

f) Criação das tabelas finais: Salvamento das quatro tabelas finais e limpas (`rooms`, `hosts`, `reviews`, `amenities`) no Unity Catalog do Databricks em formato Delta Lake.

g) Modelagem e Visualização (Power BI):

- Conexão com as tabelas do Unity Catalog via conector Azure Databricks (modo Importar).
- Criação do modelo de dados com relacionamentos entre as tabelas.
- Criação de colunas calculadas DAX para segmentação (`Stay Duration Type`, `MinimumNightsGroup`, `Occupancy Level`, `Future Demand Indicator`, `Years as Host`, `Rate Groups`, `License Status`).
- Criação de medidas DAX chave (`Avg Est Annual Occupancy`, `Avg Est Annual Income`, `Occupancy rate`, `Total Est Annual Income`, `Avg Est Annual Income PER HOST`, `Median Est Annual Income PER HOST`, `Avg Listings per Host`, etc.).

- Desenvolvimento de visuais interativos (KPIs, gráficos de barras/colunas/linhas combinados, mapas, matrizes, slicers) para explorar as hipóteses e responder à pergunta central, com foco no segmento de alta performance.

4. Insights e Recomendações

Esta seção apresenta os principais achados da análise, focando em identificar fatores que influenciam a receita das acomodações do Airbnb em Nova York, especialmente no segmento de alta performance (definido por alta ocupação estimada ≥ ~200 dias e alta demanda futura indicada por disponibilidade ≤ 90 dias).

As recomendações visam derivar estratégias para aumentar a receita da plataforma. A análise foi dividida em dois eixos principais: fatores relacionados ao anfitrião e fatores relacionados à acomodação.

4.1. Principais Insights da Análise

A análise do segmento revelou os seguintes fatores determinantes para a receita do Airbnb Nova York:

- **Localização Percebida como Fator Chave:** A nota média de reviews de localização tem uma correlação positiva com a taxa de ocupação histórica nos bairros. Bairros onde os hóspedes percebem como melhor (notas mais altas) tendem a ficar mais ocupados. Essa relação é ainda mais evidente para o segmento de estadias curtas (<30 dias). Acomodações de curta duração com notas de localização altas apresentam taxas de ocupação altíssimas. Isso sugere que turistas de curta duração, que muitas vezes não conhecem a cidade, priorizam e estão dispostos a reservar acomodações com localização comprovadamente boa. Bairros como Midtown, Financial District, Soho, Theater District e Murray Hill se destacam pela combinação de alta ocupação, alta demanda e alto custo por noite.

- **Dinâmica de Preço vs. Ocupação:** Observa-se uma relação inversa entre o preço médio diário e a taxa de ocupação; bairros com preços médios mais baixos tendem a ter ocupação maior. Isso sugere que os hóspedes são sensíveis a preço.

- **Alta rentabilidade das estadias de curta duração:** Apesar das restrições regulatórias na cidade de NY e de serem minoria em volume, as acomodações de curta duração (<30 dias) geraram a maior receita total estimada (aprox. $131M vs $110M para longa duração).

- **Perfil de Acomodação por Demanda (Tipo e Tamanho):** O tipo de acomodação preferido varia por bairro, refletindo diferentes perfis de demanda. Por exemplo, no Theater District, existe um foco em quartos privados para 2-3 pessoas, enquanto no Financial District predominam apartamentos inteiros. "Entire home/apt" geralmente comanda preços médios mais altos em relação à "Private room".

- **Otimização de Preço baseada em benchmarks locais:** Existem regiões (ex: Battery Park City, Arverne) onde acomodações com alta taxa de ocupação já conseguem praticar preços mais elevados.

- **Impacto (limitado) das Comodidades:** A análise não revelou uma correlação clara entre comodidades específicas e um prêmio significativo no preço. O valor agregado parece estar muito mais associado à localização e fatores estruturais (ex: qualidade/luxo do imóvel).

- **O Status de Superhost (SH) gera um “prêmio de preço”:** No segmento de alta performance, onde a ocupação já é similarmente alta para todos, a principal vantagem do status SH é financeira. Superhosts cobram diárias mais altas e, consequentemente, geram uma receita mediana maior, o que pode ser justificado pela sua maior visibilidade e confiança que o selo oferece, justificando um "prêmio de preço" percebido pelo cliente**.**

- **O tamanho do portfólio impulsiona a receita total:** A receita *média* por host é fortemente influenciada pelo número de propriedades gerenciadas. Observou-se que alguns Não-Superhosts (NSH) com portfólios extensos

(especialmente em Manhattan) superam a receita média de SH, embora percam na precificação média por diária. Isso é justificado pelo fato de que alguns NSH de sucesso possuem diversas propriedades alugadas, elevando assim a sua receita média total, obtendo receitas maiores que SH.

- **O ciclo de vida do anfitrião impacta a receita:** A performance da receita não aumenta linearmente com o tempo de plataforma. A receita média atinge um pico quando o anfitrião possui entre 3 e 5 anos de experiência. Anfitriões mais antigos (5+ anos) podem apresentar estagnação ou necessidade de reengajamento.

- **Métricas de Responsividade têm baixo impacto:** Para anfitriões que já estão no grupo de alta performance, taxas de resposta ou aceitação elevadas não se traduzem em ganhos adicionais significativos. O impacto dessas métricas (a taxa de resposta de 90% e a avaliação ≥4.8) está em garantir a elegibilidade ao status Superhost, e não em otimizar uma performance que já é alta.

4.2. Recomendações estratégicas

Com base nos insights identificados, as seguintes recomendações são propostas para otimizar a receita dos anfitriões no segmento de alta performance e, consequentemente, da plataforma.

Gestão de Anfitrião: Capitalizar o Status Superhost e Gerenciar o Ciclo de Vida

1. **Criar um Funil de Conversão para Superhost:** Identificar ativamente os Não-Superhosts (NSH) de alta performance e diagnosticar quais critérios específicos eles não atendem (ex: avaliação 4.75, taxa de resposta <90%). Oferecer suporte customizado e direcionado para atingir os critérios faltantes. O marketing para este grupo deve focar nos benefícios financeiros (o "prêmio de preço" dos SH na mesma área) e operacionais (suporte dedicado) do status.

2. **Reforçar o Valor de Manutenção do Status Superhost:** Para SHs existentes, comunicar continuamente que manter os altos padrões (avaliação ≥4.8, cancelamento <1%) é o que sustenta sua capacidade de cobrar diárias mais altas e manter a visibilidade, mesmo em um mercado competitivo. O risco a ser comunicado não é a perda de ocupação (que já é alta), mas a perda do status e do "prêmio de preço".

3. **Implementar Programas de Gestão de Ciclo de Vida do Host:** Reconhecer que a performance atinge um pico entre 3-5 anos. Criar programas de reengajamento para anfitriões veteranos (5+ anos), focados em otimização contínua e manutenção dos padrões SH. Para novos anfitriões (<3 anos), apoiar o crescimento dentro da plataforma, com programas de incentivos em tarifas para aumentar a quantidade de hospedagens, acelerar o aprendizado para maximizar seu potencial de receita mais rapidamente.

Precificação Inteligente: Otimização Baseada em Localização e Benchmarks

1. **Focar a Otimização de Preço em Localização, não em Comodidades:** Direcionar os anfitriões deste segmento a otimizar o preço com base em localização, benchmarks competitivos e sazonalidade. Desencorajar investimentos em comodidades adicionais cujo Retorno sobre o Investimento (ROI) não é claramente comprovado, priorizando a qualidade percebida e a localização.

2. **Implementar Recomendações Ativas de Benchmark (Preço Ótimo):** A estratégia não é apenas "aumentar o preço", mas encontrar o "ponto ótimo" que maximize a receita total (Preço x Ocupação). Identificar proativamente anfitriões em regiões de alta performance (ex: Battery Park City, Arverne) que possuem alta ocupação e características estruturais similares, mas que praticam preços abaixo da média local. Sugerir ativamente o aumento da diária para alinhá-los ao potencial máximo daquela região quando houver potencial demanda.

3. **Aplicar Precificação Premium para Segmentos de Alto Valor:** Capitalizar sobre os segmentos mais rentáveis. Usar a precificação inteligente para sugerir preços "premium" específicos para acomodações de curta duração (<30 dias) que sejam licenciadas e possuam notas altas de localização, dado que este público prioriza e paga mais por essa característica.

Marketing e Produto: Direcionar Demanda para Nichos de Alto Valor

1. **Criar Campanhas de Marketing Segmentadas por Localização e Nicho:** Substituir campanhas genéricas por marketing direcionado que explore o perfil de demanda de cada bairro. Criar campanhas (e-mail, *features* no site) destacando perfis como "Ideal para trabalho próximo ao Financial District" ou "Acomodações com melhor custo-benefício próximas aos teatros". Para turistas de curta duração, focar em campanhas que destaquem as notas de localização e proximidades (ou fácil acesso) a pontos turísticos.

2. **Marketing direcionado para hospedagens de alta ocupação com baixa demanda:** Ao considerar seu potencial devido à sua ocupação nos últimos 12 meses, criar uma campanha visando aumentar a ocupação desses lugares: Colocar no topo da lista, criar um selo de “recomendado” que fique visível na lista inicial caso ele seja um superhost (para alinhar com a ideia de incentivo para adquirir esse status)

3. **Desenvolver e Testar *Features* de Produto Focadas em Localização:** Reforçar visualmente o principal diferencial (localização) na plataforma. Testar uma nova feature (ex: um selo ou tag "Fácil Acesso às Atrações") para acomodações com notas de localização > 4.8. Se for observado um aumento na demanda (visitas/reservas) para essas acomodações, a plataforma pode proativamente recomendar um aumento de preço ao anfitrião.

4. **Proteger Estrategicamente o Segmento de Curta Duração:** Mitigar o risco regulatório sobre o segmento de maior receita total. Dado o seu alto valor de receita, apoiar ativamente os anfitriões deste segmento para que obtenham e mantenham suas licenças (status license) e construir uma relação

harmoniosa com a prefeitura de Nova York para evitar futuros embates, é fundamental priorizar a proteção desse nicho para garantir e ampliar essa receita.

5. **Diversificação de mercado:** Analisar os bairros com perfil de longa duração para incentivar esse tipo de estadia, estudar o mercado de aluguéis de longo prazo da região para entrar nele como um protagonista. O principal incentivo pode ser diminuição da taxa do airbnb e até isenção por um período determinado, fazendo com que os hosts de plataformas mais estabelecidas queiram migrar pro airbnb e, posteriormente, aumentar essas taxas quando a empresa estiver com uma parcela considerável de presença nesse mercado.

5. Aprendizados e Reflexões

O processo de análise deste projeto forneceu diversos aprendizados valiosos sobre o tratamento e a interpretação de dados. Primeiramente, destacou-se a importância fundamental de um processo de limpeza iterativa e validação constante dos dados, abordando questões como a diferenciação entre "N/A" e valores nulos reais, e a identificação e tratamento sistemático de outliers, que podem distorcer significativamente as métricas quando ignorados.

Compreendemos também o poder da segmentação para revelar insights contextuais profundos. A análise geral do dataset apresentava conclusões diferentes da análise focada no segmento de alta performance, mostrando como o contexto altera a interpretação dos resultados. Nesse sentido, tornou-se evidente a diferença crucial entre média e mediana em distribuições assimétricas, sendo a mediana frequentemente a medida mais robusta para descrever o cenário "típico", tanto para taxas de host quanto para métricas de review. Isso porque, em distribuições assimétricas, a média pode levar a interpretações distorcidas por sofrer influências de dados extremos que podem elevar ou reduzir o seu valor.

Além disso, o projeto reforçou a necessidade de combinar múltiplas métricas para entender fenômenos complexos, como a receita, que é um produto direto do preço e da ocupação. Igualmente importante foi a constatação da necessidade de

buscar informações em fontes externas para contextualizar a análise. Por exemplo, a pesquisa sobre as restrições regulatórias de Nova York para hospedagens de curta duração foi essencial para interpretar corretamente a coluna `minimum_nights` e segmentar o mercado . Da mesma forma, encontrar a taxa de conversão de reviews em reservas utilizada pelo Inside Airbnb (72%, baseada em dados históricos) permitiu estimar métricas como a duração média da estadia, enriquecendo a análise. A busca por dados de localização de pontos turísticos, demonstrou o valor de integrar dados contextuais externos para entender fatores que influenciam o sucesso das hospedagens.

Finalmente, a escolha da ferramenta certa para cada etapa provou ser essencial, reconhecendo a complexidade e a adequação de cada tecnologia: PySpark para o ETL inicial e manipulações pesadas, Power Query para transformações e limpeza mais diretas dentro do ambiente de BI, e DAX para a criação de métricas dinâmicas e agregações complexas no modelo de dados.

Durante o processo, enfrentamos alguns desafios. A utilização da linguagem PySpark, por não ser uma ferramenta de domínio prévio da equipe, apresentou uma curva de aprendizado inicial. Notavelmente, a complexidade do tratamento da coluna `amenities` exigiu a estruturação de um modelo específico em PySpark para extrair, limpar e transformar a lista de comodidades (contida em uma única string complexa) em dados utilizáveis para análise. Enfrentamos desafios com a complexidade da linguagem M do Power Query para implementar lógicas estatísticas como o IQR, que posteriormente foi introduzida como uma etapa no Databricks. A construção de fórmulas DAX corretas para agregações aninhadas, como o cálculo da receita média por host, também exigiu atenção aos detalhes de contexto. Interpretar resultados, como a divergência entre média e mediana, e garantir a integridade referencial dos dados após aplicar filtros foram desafios superados através de pesquisa, testes iterativos e raciocínio lógico aplicado aos dados.

Refletindo sobre o processo, em um próximo projeto, poderíamos definir o escopo do dataset final (como os filtros de alta performance) talvez um pouco mais cedo, após as análises exploratórias iniciais, para otimizar as etapas subsequentes de limpeza. Apesar do esforço inicial de tratamento, poderíamos também ter

explorado mais a fundo a transformação e análise individual das comodidades (`amenities`), o que poderia gerar insights ainda mais granulares. Por fim, dedicar um tempo adicional à validação rigorosa dos tipos de dados logo após a leitura inicial poderia ter evitado algumas refatorações posteriores. No geral, o projeto foi uma experiência rica em aprendizados práticos sobre o ciclo completo de análise de dados, incluindo a integração crucial entre os dados brutos e o contexto externo.

6. Referências

- Fontes de Dados:

DGomonov. (Data Publisher). *New York City Airbnb Open Data*. Kaggle Dataset. Acessado em 14 out. 2025, disponível em: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Inside Airbnb. *Get the Data - New York City*. Acessado em 14 out. 2025, disponível em: http://insideairbnb.com/get-the-data/.

- Dicionário de Dados:

Dicionário de Dados - Airbnb NYC Analysis. Google Sheets. Acessado em 14 out. 2025, disponível em: https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit?usp=sharing

- Materiais de Apoio:

NYC Office of Special Enforcement (OSE). *Short-Term Rentals Information*. Disponível

em:https://www.nyc.gov/site/specialenforcement/registration-law/registration-rules-and-laws.page.

San Francisco Planning Department. (2015, April 23). *Executive Summary: Amendments Relating to Short-Term Rentals (Case Nos. 2014-001033PCA, 2015-003861PCA, and 2015-004765PCA)*. Staff Report to the Planning Commission Hearing. Acessado 24 out. 2025, disponível em: https://commissions.sfplanning.org/cpcpackets/2014-001033PCA.pdf.

- Ferramentas:

Databricks (Plataforma de Análise de Dados)

PySpark (Processamento de Dados)

Microsoft Power BI (Visualização de Dados e Modelagem)