

[Upgrade](#)[Open in app](#)

Gustavo Santos

[Follow](#)

Jul 23, 2020 · 5 min read



Save



## Insights do dataset Airbnb NYC

O Airbnb é uma empresa de San Francisco, CA nos EUA. Seu principal negócio é o mercado de aluguel, onde eles atuam no modelo PaaS (Plataforma como Serviço), oferecendo uma plataforma na qual os usuários podem listar suas casas, apartamentos ou até um quarto individual para aluguel.

Esse é um modelo de negócios bem-sucedido que vem crescendo em muitos segmentos (transporte, serviços de streaming etc.) e gerando dados que podem nos ajudar a responder a muitas perguntas.

Neste projeto, usei a linguagem Python para explorar os “Dados abertos do Airbnb da cidade de Nova York” ([link para download](#)) e extrair o maior número possível de informações.

Algumas das perguntas para as quais eu estava procurando resposta são:

- O que podemos aprender sobre as diferentes áreas?

- Quais hosts são os hosts mais populares e por que?

- Existe alguma diferença notável de tráfego entre diferentes áreas e qual poderia ser a razão disso?

## O conjunto de dados

Este conjunto de dados possui 16 colunas (a saber: `'id'`, `'name'`, `'host_id'`, `'host_name'`, `'neighbourhood_group'`, `'neighbourhood_group'`, `'neighbourhood_group'`, `'neighbourhood'`, `'latitude'`, `'longitude'`, `'room_type'`, `'price'`, `'minimum_nights'`, `'number_of_reviews'`, `'last_review'`, `'reviews_per_month'`, `'protected_host_listings_count'`, `'Availability_365'`) e 48.895 observações.

Os insights iniciais extraídos após o uso da função `.describe()` foram:

- Nome do host 'Michael' é o primeiro, com 417 aparições;
- A maioria das propriedades fica na ilha de Manhattan (21,6k de 48,8k = 44%);
- Mais de 50% dos aluguéis são para "Casa ou apto", não apenas um quarto;
- Em média, a taxa é de US \$ 152 e isso não parece variar muito, já que 75% das operações são inferiores a US \$ 175;
- A média de estadia é de 7 noites. As pessoas aparentemente gostam de ficar até uma semana inteira em Nova York;


Em uma fase de exploração, é interessante criar histogramas, para que você possa ver como os dados estão distribuídos.

### **Tempo, Custo e Onde**

O próximo passo foi descobrir quanto tempo as pessoas ficam, quanto pagam e os locais mais procurados.

Como resultado, descobri que as pessoas procuram **uma semana de estadia** em Nova York, mas isso pode variar de 2 a 7 dias. Além disso, há uma boa parte dos aluguéis na média de 30 dias. Com uma exploração mais aprofundada, pude concluir que os 10 principais hosts concentram seus aluguéis para negócios e não para turismo, mantendo assim suas propriedades alugadas por uma média de 30 dias de duração.

A despesa média está entre US \$ 120 e US \$ 150 por noite, o que era um bom negócio em comparação com hotéis em locais semelhantes em 2019. Além disso, a partir dessa variável também é possível ver como as viagens de negócios influenciam os preços em Nova York, pois havia algumas observações com preços nas faixas de US \$ 200 e até US \$ 300 +.

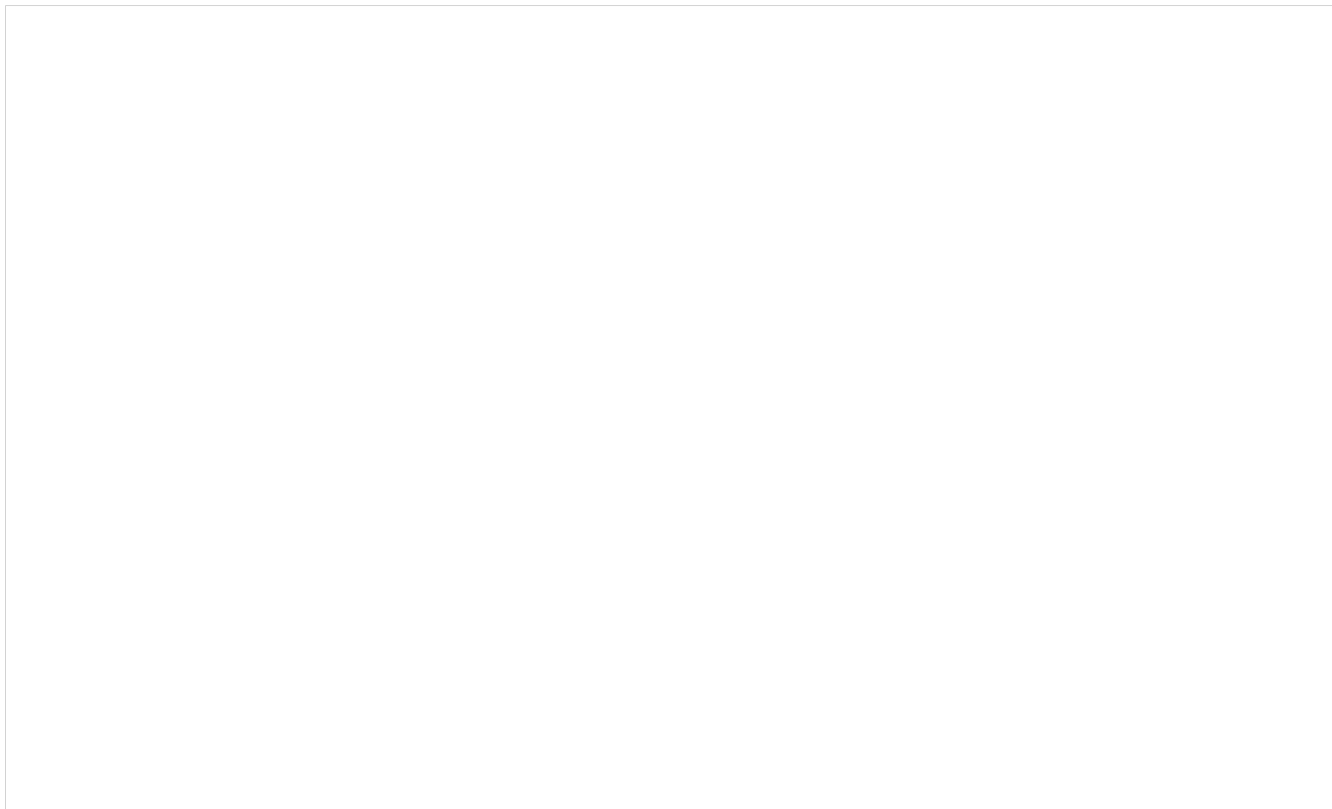


Veja como o gráfico muda com a limpeza dos outliers. À direita, o gráfico dos dados limpos.

Além disso, em relação à localização, 85% das listagens são de Manhattan (44%) e Brooklyn (41%, provavelmente devido à sua proximidade com a ilha). E isso não é uma surpresa, na verdade. Manhattan é o coração econômico da cidade, concentrando os arranha-céus com seus inúmeros escritórios e famosas atrações turísticas de classe mundial, como *Times Square*, *5th Ave* ou *Central Park*.

Milhões de pessoas vão à Nova York todos os anos para fazer negócios ou lazer e pretendem ficar o mais próximo possível de Manhattan para aproveitar o local, evitando longas viagens de metrô ou carro, consequentemente economizando tempo. Naturalmente, seguindo a regra de oferta versus demanda, é provável que os preços sejam mais altos nessa região.

Aqui está um *HeatMap* criado usando o pacote Folium Map para demonstrar onde estão os imóveis e os preços (sendo o vermelho os mais caros).



HeatMap: mais vermelho, mais alto o preço e mais concentrado o número de listagens.

### **Os Top 10 hosts**

Depois de finalizar as visualizações de todo o conjunto de dados, criei esse subconjunto com os 10 hosts mais populares em número de aluguéis para ver o que eles tinham diferente dos outros que poderiam torná-los bem-sucedidos.

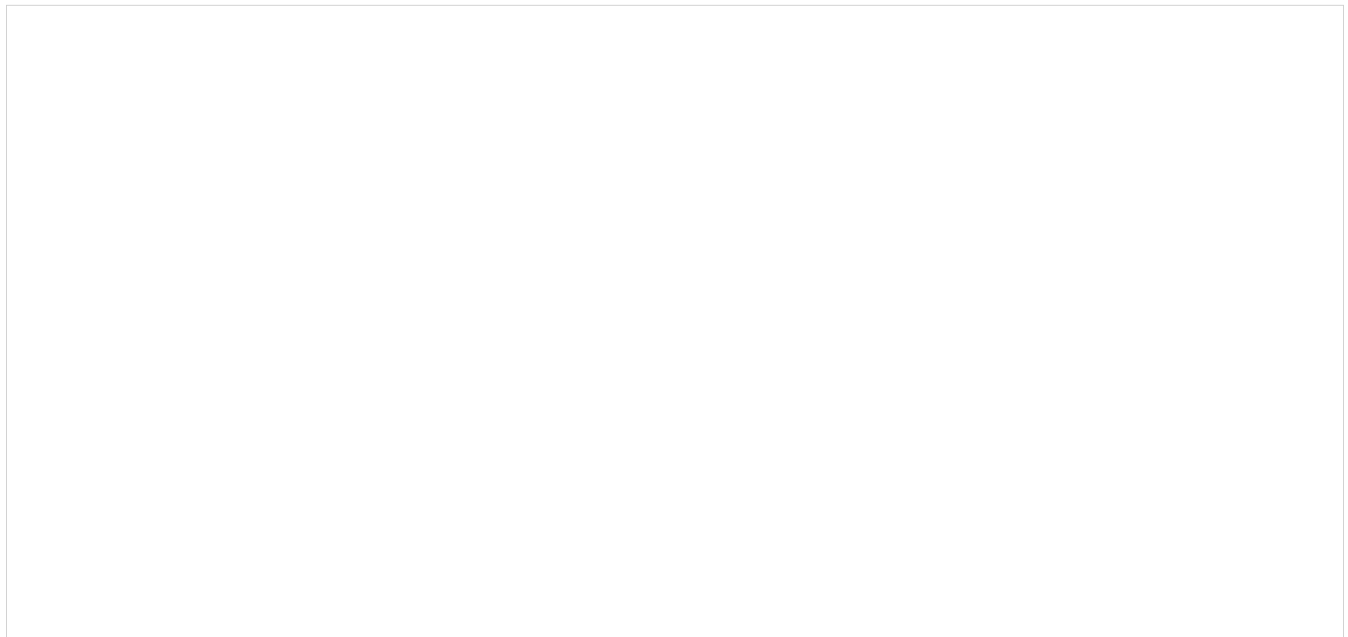
É interessante notar que os 10 principais hosts representam 2,6% do total de aluguéis, em uma amostra de 37.500 hosts. Isso é um número significativo!

**Localização:** o primeiro e óbvio motivo foi encontrado no gráfico onde pude ver que todos eles tinham propriedades localizadas em Manhattan e / ou Brooklyn. Faz muito sentido conectar-se às facilidades daquele local.

**Esse grupo cobra mais caro do que os preços médios.** Novamente, isso é uma consequência de muitas opções que eles têm no terreno mais caro de Nova York. A localização é geralmente um fator-chave quando se trata do mercado

imobiliário.

**Eles provavelmente se concentram nos aluguéis para empresas em detrimento dos aluguéis de turismo.** Esta conclusão vem do gráfico que mostra os top 10 com uma duração média de locação de 30 dias contra 7 dias, se considerarmos o conjunto de dados inteiro. 30 dias se parecem muito mais com aluguel de empresas do que com férias. Além disso, após uma rápida verificação na Internet, vi que *Sonder*, *Vida*, *Blueground* são todas empresas do mercado de aluguel em Nova York, por isso têm uma equipe e uma estrutura de suporte para seus clientes (recursos normalmente exigidos por outras empresas).



### **Processamento de Linguagem Natural (PLN)**

Para concluir o projeto, realizei uma PLN usando o pacote *nlkt* para extrair alguns insights das revisões feitas pelos usuários. Criei uma nuvem de palavras e uma contagem de frequência de palavras para ver as palavras mais frequentes. *Quarto*, *aconchegante*, *Brooklyn*, *Manhattan*, *privado*, *Studio* estavam entre as principais palavras usadas.

Depois, mudei para a análise de Bigramas e Trigramas — *duas / três palavras usadas juntas e em sequência dentro de um texto* — pois isso nos dá uma melhor compreensão de como as palavras são usadas em um contexto. O resultado mostrou que os usuários estão alugando quartos privativos, apartamentos de 1 ou 2 quartos. Eles gostam de ficar no 'Upper East Side' perto da 5th Avenue e Central Park para aproveitar essa área bem localizada. Comentários que mencionam qualidades subjetivas, como *aconchegante* ou *casa longe de casa*, apareceram entre os mais frequentes, demonstrando que os clientes estão preocupados com a limpeza e a manutenção do local.

• • •

## **Concluindo**

A ciência de dados é mais do que previsões. Trata-se de extrair boas informações e conhecimentos dos dados.

Neste projeto, eu pude executar muitas tarefas comumente usadas para exploração e visualização de dados, extraindo informações do conjunto de dados Airbnb para ajudar as empresas a tomarem decisões melhores e baseadas em dados.

Para verificar o Jupyter Notebook, acesse meu repositório do GitHub:  
<https://github.com/gurezende/NYC-Airbnb>.

```
if data:
    data.science()
```

*Gus*



---

## Get an email whenever Gustavo Santos publishes.

Emails will be sent to narendrayadav1610@gmail.com. [Not you?](#)



Subscribe