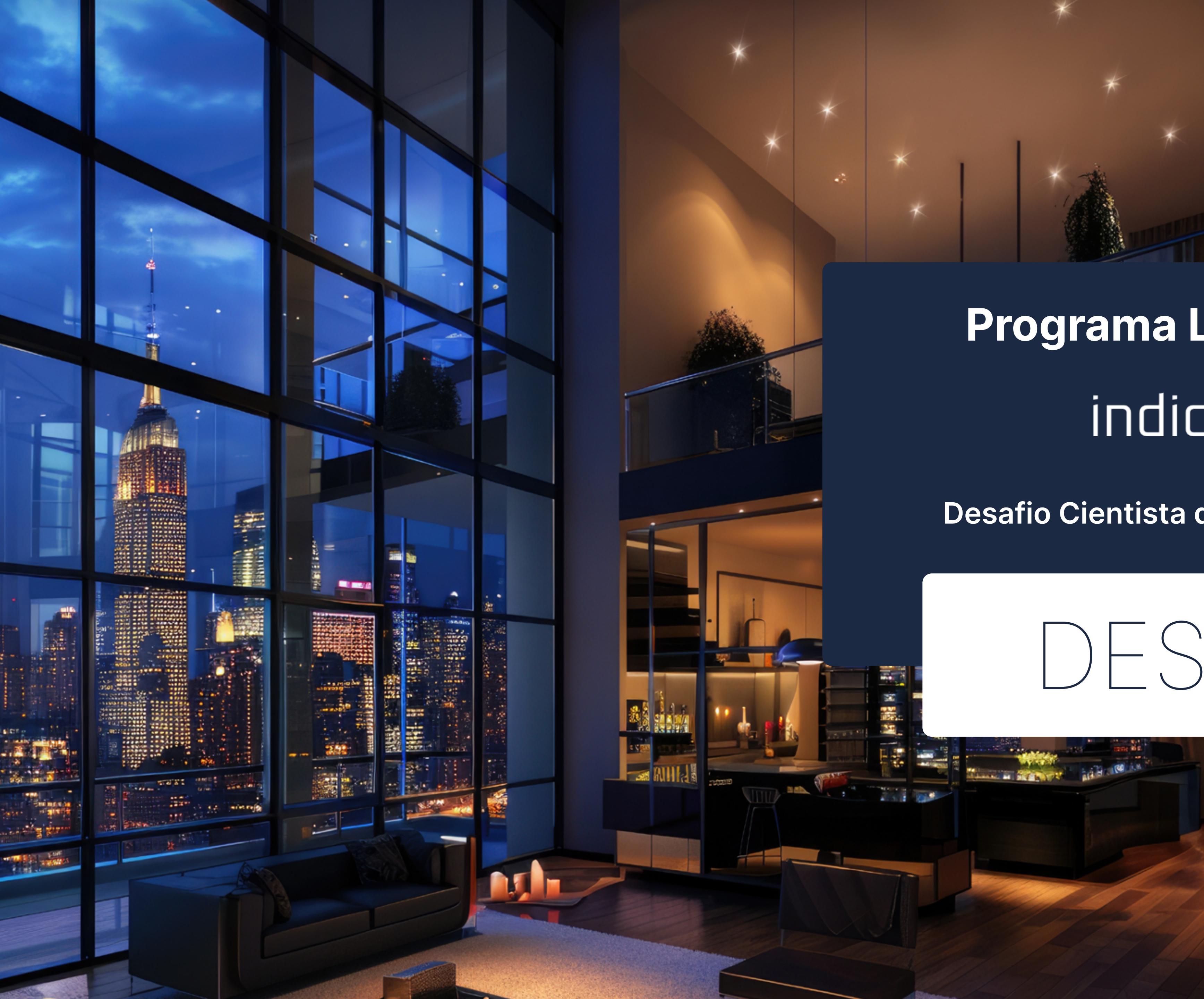


Programa LIGHTHOUSE



Desafio Cientista de Dados - NY Rental

Luiz Augusto Soutes



Programa LIGHTHOUSE



Desafio Cientista de Dados - NY Rental

DESAFIO

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

OBJETIVO DESTE DESAFIO

Desenvolver um modelo de previsão de preços a partir do dataset oferecido, e avaliar tal modelo utilizando as métricas de avaliação que mais fazem sentido para o problema.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ENTREGAS ESPERADAS

1. Análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas.
2. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?
3. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?
4. Existe algum padrão no texto do nome do local para lugares de mais alto valor?
5. Explique como você faria a previsão do preço a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?
6. Sugestão de preço para uma nova propriedade

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

INFORMAÇÕES GERAIS

Nesta apresentação não irei mencionar todos os códigos, visto que o objetivo aqui é mostrar o porque e o resultado das análises.

Todos os códigos podem ser vistos no arquivo: docs > rental_price_ny_EDA.ipynb

Por questão estética, os gráficos utilizados nesta apresentação foram gerados com a biblioteca Plotly, mas devido a uma incompatibilidade da mesma com a IDE utilizada, para exibir o gráfico inline no arquivo ipynb, também foi utilizada a biblioteca Matplotlib.

Para melhor navegabilidade, utilize o Adobe Acrobat e navegue através do menu lateral.

A photograph of the Manhattan Bridge in New York City, viewed from a street level between two red brick buildings. The bridge's steel towers and cables are prominent against a blue sky with white clouds.

Programa LIGHTHOUSE



Desafio Cientista de Dados - NY Rental

A large, white, rounded rectangular box centered on the right side of the slide, partially overlapping the dark blue header area.

ETAPAS

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ETAPAS DO PROJETO



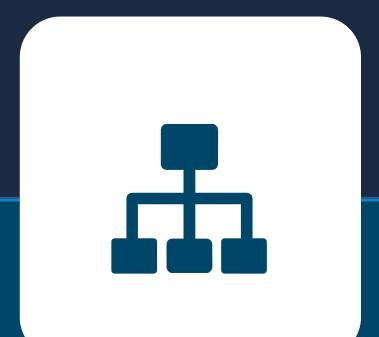
Entendimento
do negócio



Tratamento
dos dados



Análise
exploratória
de dados



Análise
preditiva



Teste de
performance
do modelo
com predição

Estas duas etapas estão no arquivo
`rental_price_ny_model.ipynb`



Programa LIGHTHOUSE



Desafio Cientista de Dados - NY Rental

NEGÓCIO

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ENTENDIMENTO DO NEGÓCIO

Nova Iorque é uma das cidades mais visitadas no mundo.

Seus visitantes, podem ir a trabalho, mas a grande maioria vai como turista.

Uma opção de hospedagem são as propriedades de locação de curto prazo.

Os preços dessas propriedades variam muito, pois estes podem depender da sua localização (seja num bairro mais valorizado ou por estar próximos a ponto turísticos), tipo do quarto, entre outros.

Veremos um pouco mais sobre isso nos gráficos a seguir.



Programa LIGHTHOUSE



Desafio Cientista de Dados - NY Rental

DADOS

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• BOAS PRÁTICAS

Este projeto segue as orientações da PEP 8, que é um conjunto de recomendações onde os códigos sejam mais legíveis e consistentes.

De acordo com essas orientações todos os códigos estão comentados e com o mesmo padrão de nomenclatura e indentação.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• BIBLIOTECAS UTILIZADAS

Para este projeto serão utilizadas as seguintes bibliotecas e suas respectivas funções

CÁLCULOS E ESTRUTURAS

Numpy
Pandas

VISUALIZAÇÃO DOS DADOS

Matplotlib.pyplot
Seaborn
Plotly

MODELAGEM

Sklearn
XG Boost

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• DATASET

```
df = pd.read_csv("../docs/teste_indicium_precificacao.csv")
df.head()
```

	id	nome	host_id	host_name	bairro_group	bairro	latitude	longitude	room_type	price
0	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225
1	3647	THE VILLAGE OF HARLEM....NEW YORK!	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150
2	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89
3	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80
4	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• DATASET

O dataset fornecido foi o arquivo teste_indicium_precificacao.csv

Este contém 48.894 entradas distribuídas em 16 variáveis.

id – Atua como uma chave exclusiva para cada anúncio nos dados do aplicativo

nome - Representa o nome do anúncio

host_id - número de identificação do anfitrião

host_name - Nome do anfitrião

bairro_group - Principais regiões da cidade

bairro - Bairros

latitude - Latitude da propriedade

longitude - Longitude da propriedade

room_type - Tipo de acomodação

price - Preço por uma noite

minimo_noites - Quantidade mínima de noites para reservar

numero_de_reviews - Número de avaliações recebidas

ultima_review - Data da última avaliação

reviews_por_mes - Quantidade de avaliações por mês

calculado_host_listings_count - Número de propriedades disponíveis no Airbnb pertencentes ao anfitrião

disponibilidade_365 - Número de dias disponíveis dentro de 365 dias

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• PADRONIZAÇÃO

Ao visualizar o conjunto de dados do dataset, identifiquei que os nomes das variáveis precisam de um ajuste para seguir um padrão de idioma. Atualmente estão utilizando inglês e português, como os dados se referem a imóveis em Nova Iorque e as variáveis que estão

Para isso foi utilizada a função **rename columns** e as variáveis ficaram conforme a lista ao lado.

id
name
host_name
neighbourhood_group
neighbourhood
latitude
longitude
room_type
price
minimum_nights
number_of_reviews
last_review
reviews_per_month
calculated_host_listings_count
availability_365

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• PADRONIZAÇÃO

Além de padronizar o nome das variáveis, pude perceber que os registros na variável nome estavam fora de padrão.

Então utilizei a função **str.title()** para que cada palavra comece com letra maiúscula.

Ao lado uma amostra de como ficou após a padronização.

```
df['name'] = df['name'].str.title()
```

```
print(df['name'].head())
```

```
0 Skylit Midtown Castle
1 The Village Of Harlem....New York !
2 Cozy Entire Floor Of Brownstone
3 Entire Apt: Spacious Studio/Loft By Central Park
4 Large Cozy 1 Br Apartment In Midtown East
Name: name, dtype: object
```

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• DADOS FALTANTES

Utilizando as funções **isnull()** e **sum()** foi constatada a presença de aprox 20% de dados faltantes nas variáveis `last_review` e `reviews_per_month`.

Estas se referem as avaliações dos usuários nas propriedades. Como este projeto tem o objetivo de sugerir o valor do aluguel de uma nova propriedade, as avaliações referentes as demais não são relevantes. A decisão tomada, foi a exclusão destas do dataframe.

Já as variáveis `host_name` e `name` contém poucos dados faltantes em relação ao total de entradas, sendo assim, deixei as mesmas por enquanto e futuramente, e se preciso, farei o tratamento adequado conforme a necessidade.

As demais variáveis foram conferidas e estão dentro do esperado.

```
A quantidade de dados nulos é:  
last_review           10052  
reviews_per_month     10052  
host_name              21  
name                   16  
neighbourhood_group      0  
neighbourhood          0  
id                      0  
host_id                 0  
longitude                0  
latitude                  0  
room_type                  0  
price                     0  
number_of_reviews         0  
minimum_nights            0  
calculated_host_listings_count 0  
availability_365           0  
dtype: int64.
```

```
Já o percentual de dados nulos é:  
last_review             20.56  
reviews_per_month        20.56  
host_name                0.04  
name                     0.03  
neighbourhood_group      0.00  
...  
minimum_nights             0.00  
calculated_host_listings_count 0.00  
availability_365           0.00
```

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• OUTLIERS

Com o método **describe()** temos conhecimento de alguns informações estatísticas, sendo elas Count (Mostra a quantidade de entradas, uma alternativa para identificar se a variável possui números faltantes), Mean (média aritmética), Std (desvio padrão), Min e max (valores mínimo e máximo), Quartis (valores que indicam a distribuição dos dados). Usei esse metodo apenas nas variáveis numéricas relevantes para o projeto.

```
df[['price', 'minimum_nights', 'calculated_host_listings_count', 'availability_365']].describe()
```

Python

	price	minimum_nights	calculated_host_listings_count	availability_365
count	48894.000000	48894.000000	48894.000000	48894.000000
mean	152.720763	7.030085	7.144005	112.776169
std	240.156625	20.510741	32.952855	131.618692
min	0.000000	1.000000	1.000000	0.000000
25%	69.000000	1.000000	1.000000	0.000000
50%	106.000000	3.000000	1.000000	45.000000
75%	175.000000	5.000000	2.000000	227.000000
max	10000.000000	1250.000000	327.000000	365.000000

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

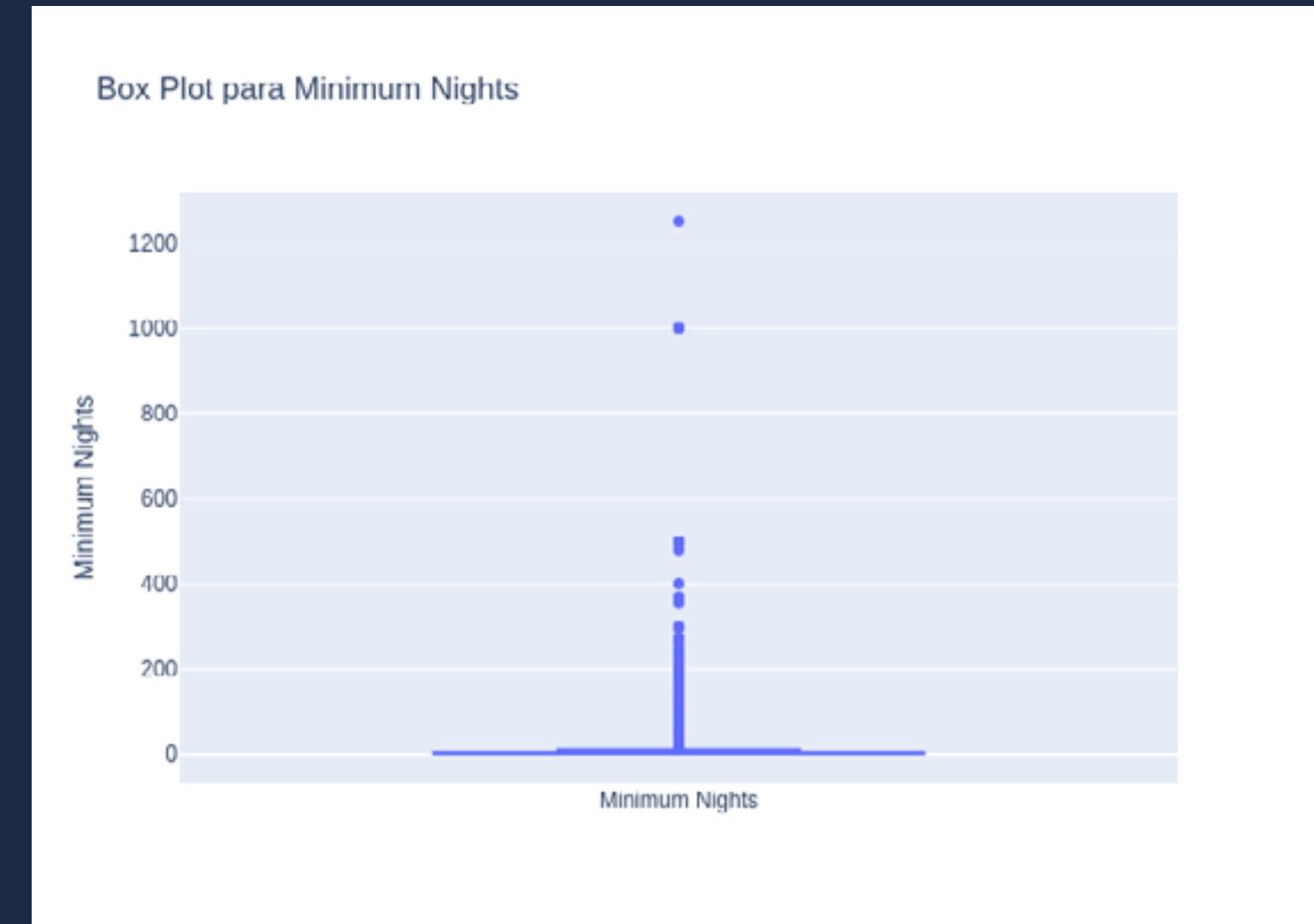
ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• OUTLIERS

Além disso utilizei gráficos do tipo Boxplot para facilitar a análise:



DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• OUTLIERS

Observando o resultado, listei alguns dados que precisam de atenção:

Na variável PRICE:

O mínimo é 0 e o máximo 10000. Como a média é aproximadamente 152 e em 75% das entradas os preços estão iguais ou abaixo de 175, os valores 0 e 10000 podem ser considerados como outliers.

Na variável MINIMUM_NIGHTS:

Pedir 1250 noites em uma propriedade temporária como minimo não é impossível, mas é muito estranho.

Na variável CALCULATED_HOST_LISTINGS_COUNT:

Aqui os dados parecem normais, exceto que temos 327 propriedades em um único usuário. Podemos considerar que estamos falando de algum profissional do ramo de locação ou até mesmo um hotel.

Na variável AVAILABILITY_365:

Temos 25% das entradas iguais a 0 e a máxima corresponde a 1 ano.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• OUTLIERS

O tratamento destes outliers se deu da seguinte maneira:

MINIMUM_NIGHTS:

Considerei neste projeto apenas as estadias de curto prazo, ou seja, com o prazo de 30 ou menos noites.

Sabendo que temos poucos mais 1,5% dos dados acima de 30 noites, limpar estes outliers é benéfico para um resultado mais preciso do modelo.

```
over30_nights = len(df[df['minimum_nights'] > 30])
over30_nights_percent = ((over30_nights) / df.shape[0]) * 100
print(f"Temos {over30_nights} registros acima de 30 noites.")
print(f"Isto corresponde a {over30_nights_percent:.2f}% dos dados")
```

Python

Temos 747 registros acima de 30 noites.
Isto corresponde a 1.53% dos dados

Sabendo que temos poucos mais 1,5% dos dados acima de 30 noites, podemos considerar que limpar este outliers é benéfico para um resultado mais preciso do modelo.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• OUTLIERS

Já em relação ao outliers na variável **Price**:

Na resultado do **describe()** observei que na variável price a média é aprox 150 e que o 4º quartil (75%) corresponde a 175. Porém foi percebi que existem alguns dados que estão muito longe disso.

Outro detalhe foi que o valor minimo constam alguns valores iguais a 0. Como 0 não corresponde a um locação válida, iremos remove-la em seguida.

O resultado apresentou que 24% dos dados estão acima do 3º Quartil (175), mas apenas pouco mais de 2% estão com valores acima de 500. Então, seguindo a mesma lógica usada em `minimum_nights`, pensando na qualidade do modelo, removi da base os dados acima de \$500 e também os valores igual a 0.

```
# Calcular o número total de propriedades
total_properties = len(df)

# Contar propriedades com preço maior que 175, 500, 1000 e 2000
above_175 = df[df['price'] > 175].shape[0]
above_500 = df[df['price'] > 500].shape[0]
above_1000 = df[df['price'] > 1000].shape[0]
above_2000 = df[df['price'] > 2000].shape[0]

# Calcular percentuais
percent_above_175 = (above_175 / total_properties) * 100
percent_above_500 = (above_500 / total_properties) * 100
percent_above_1000 = (above_1000 / total_properties) * 100
percent_above_2000 = (above_2000 / total_properties) * 100
```

Propriedades com preço acima de \$175: 11973
Percentual de propriedades acima de \$175: 24.87%

Propriedades com preço acima de \$500: 1012
Percentual de propriedades acima de \$500: 2.10%

Propriedades com preço acima de \$1000: 213
Percentual de propriedades acima de \$1000: 0.44%

Propriedades com preço acima de \$2000: 72
Percentual de propriedades acima de \$2000: 0.15%

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

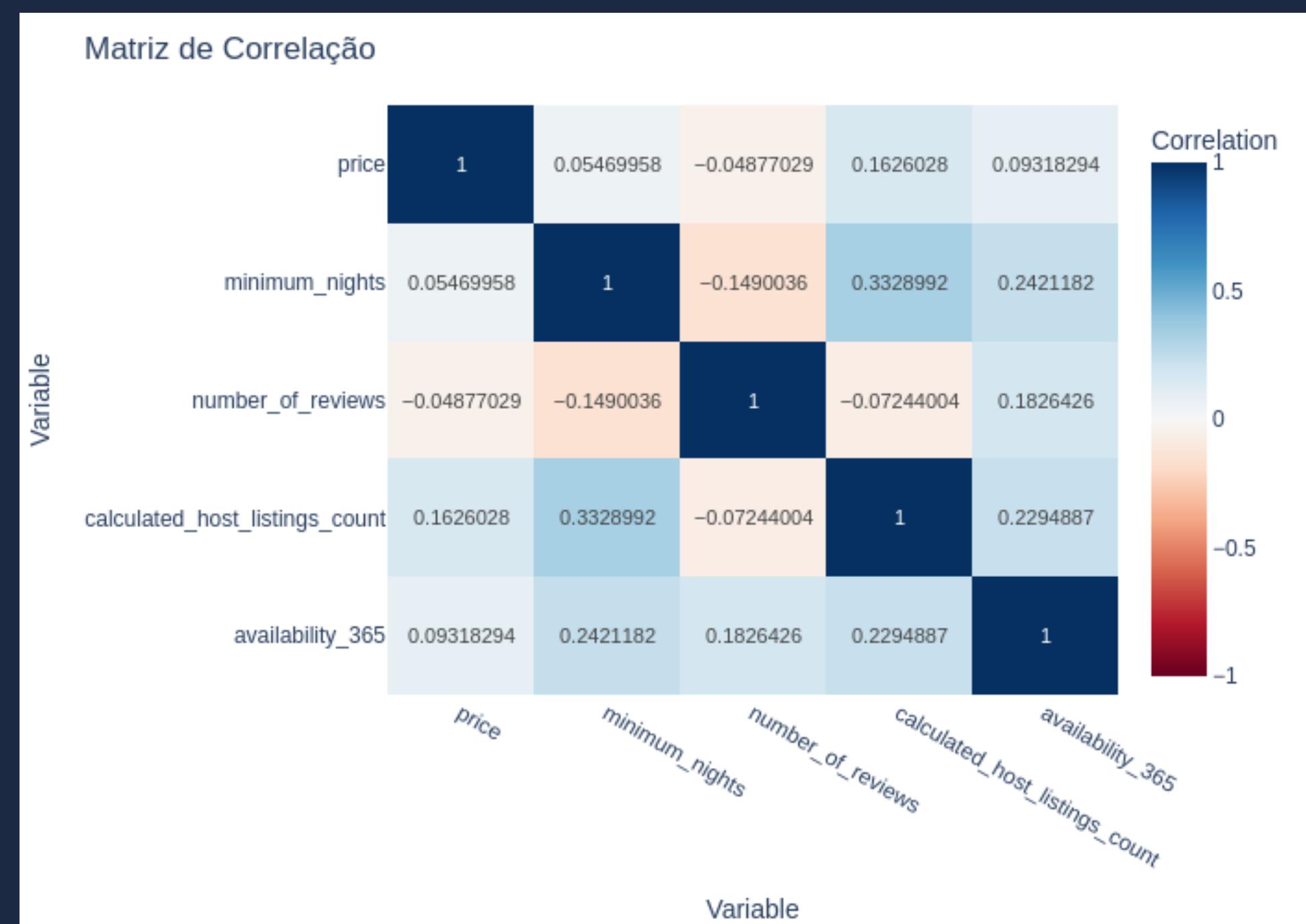
ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

TRATAMENTO DOS DADOS

• CORRELAÇÃO



Não existe nenhuma correlação significativa.



Programa LIGHTHOUSE



Desafio Cientista de Dados - NY Rental

ANÁLISES

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

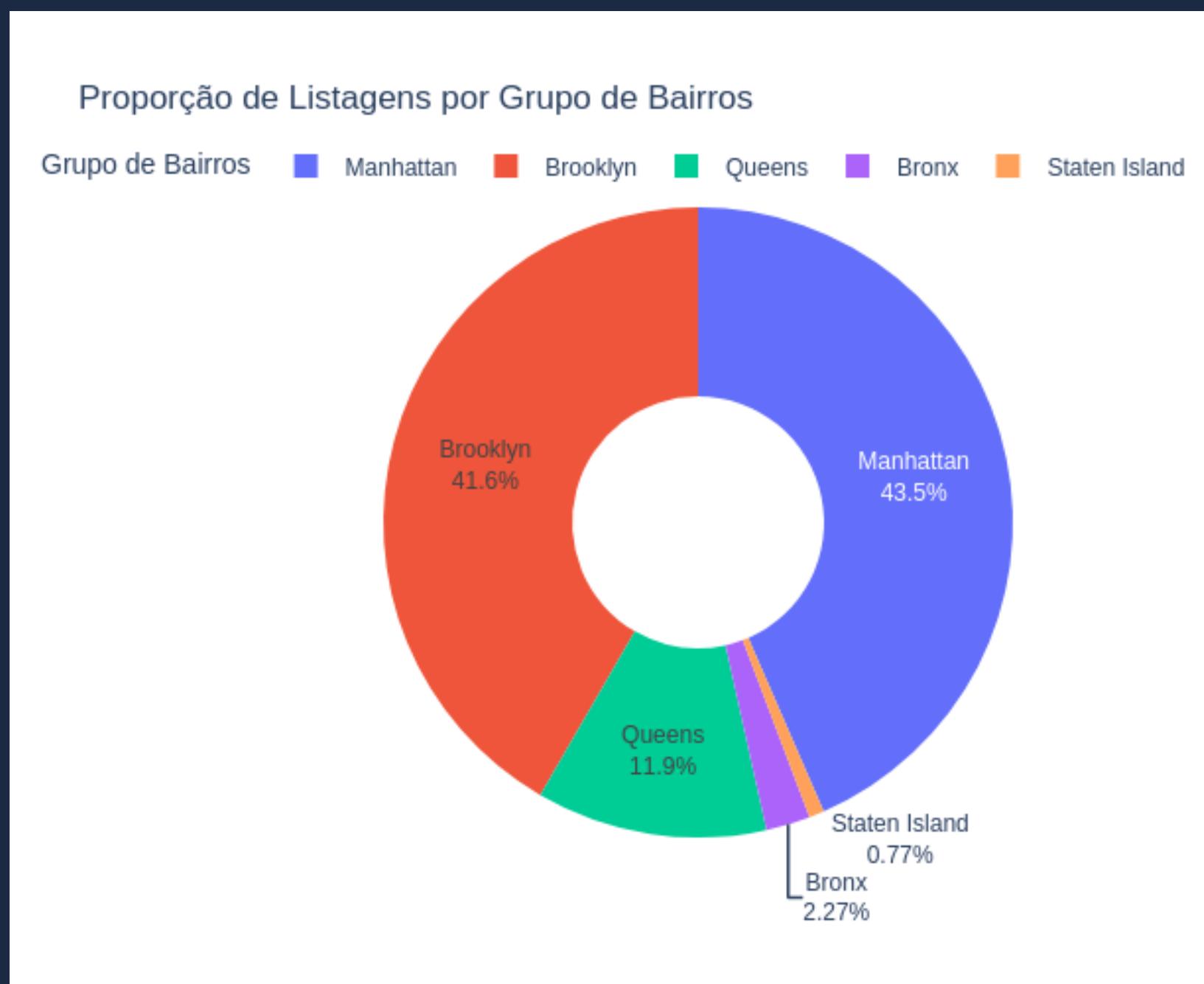
ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• GRUPO DE BAIRROS (PROPORÇÃO)



Aqui utilizei um código para a contagem dos dados da variável “neighbourhood_group” e em seguida exibir essa informação em um gráfico de rosca, que demonstra a proporção de listagens por grupo de bairros.

Neste gráfico é possível ver, em porcentagem, que temos mais de 80% das propriedades localizadas em duas regiões. Essas duas regiões estão localizadas na região central e com maior interesse da cidade.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

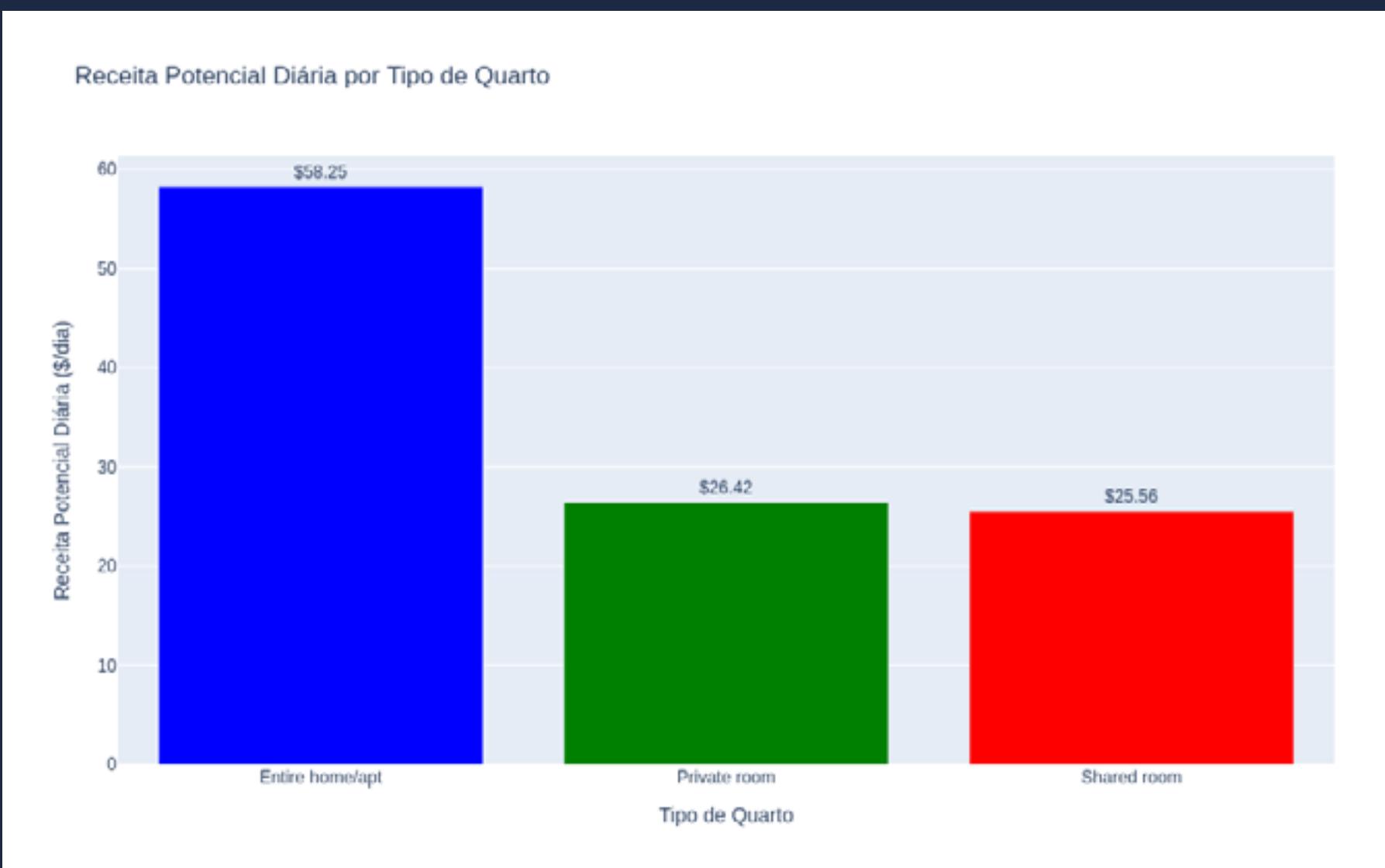
ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

- POTENCIAL RECEITA POR TIPO DE PROPRIEDADE



Podemos perceber que uma propriedade inteira tem um potencial acima do dobro das demais.

Aqui podemos começar a imaginar, será que sempre é mais rentável alugar um imóvel inteiro?

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

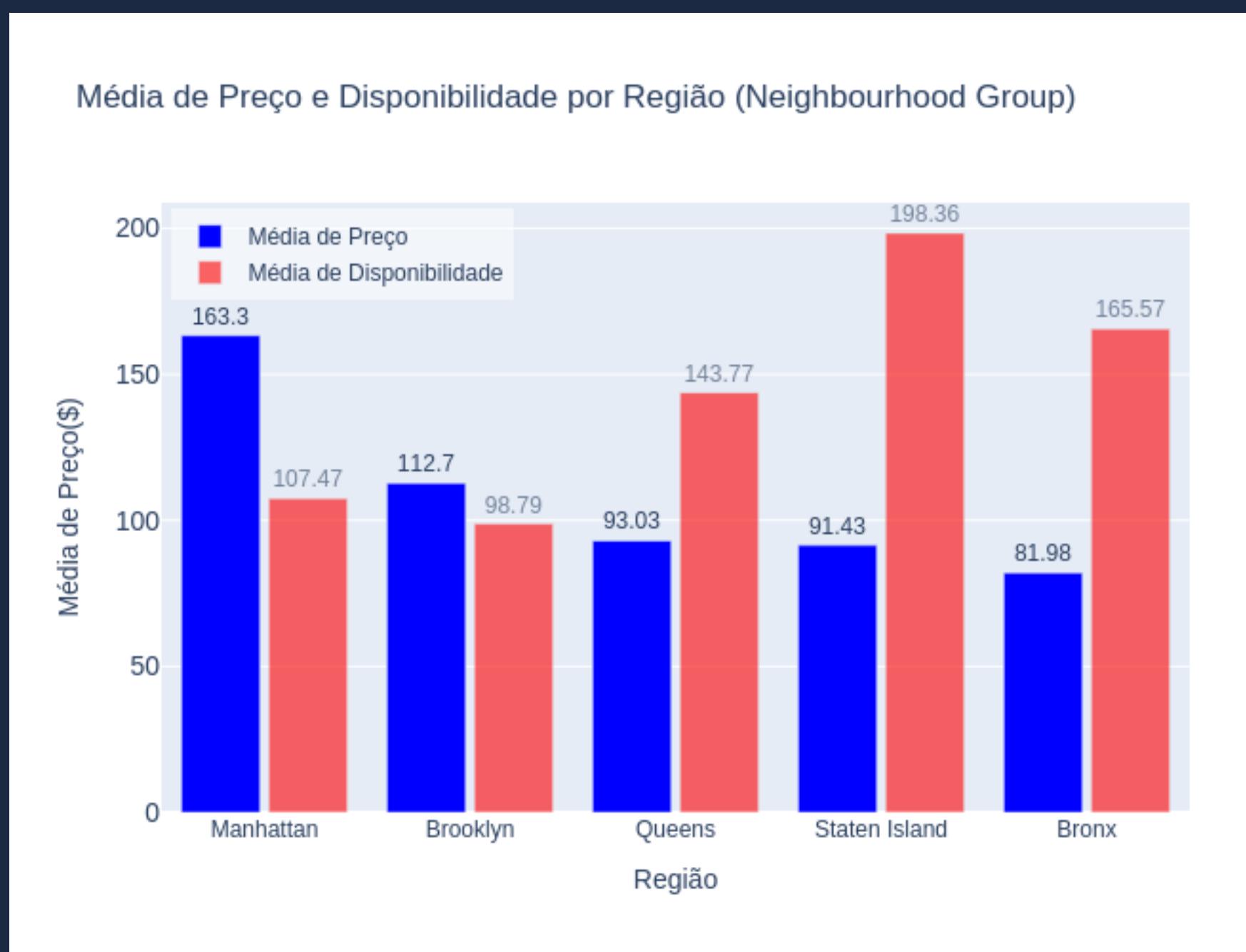
ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros

- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

- MÉDIA DE PREÇO E DISPONIBILIDADE NOS GRUPOS DE BAIRROS



Ao comparar as médias dos preços e da disponibilidade por região dos bairros, vemos que nas regiões onde tem menos propriedades a disponibilidade ao longo do ano é maior, porém a média de preço é menor.

Já nas duas regiões onde as propriedades apresentam uma menor disponibilidade, a média de preço supera as demais regiões.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

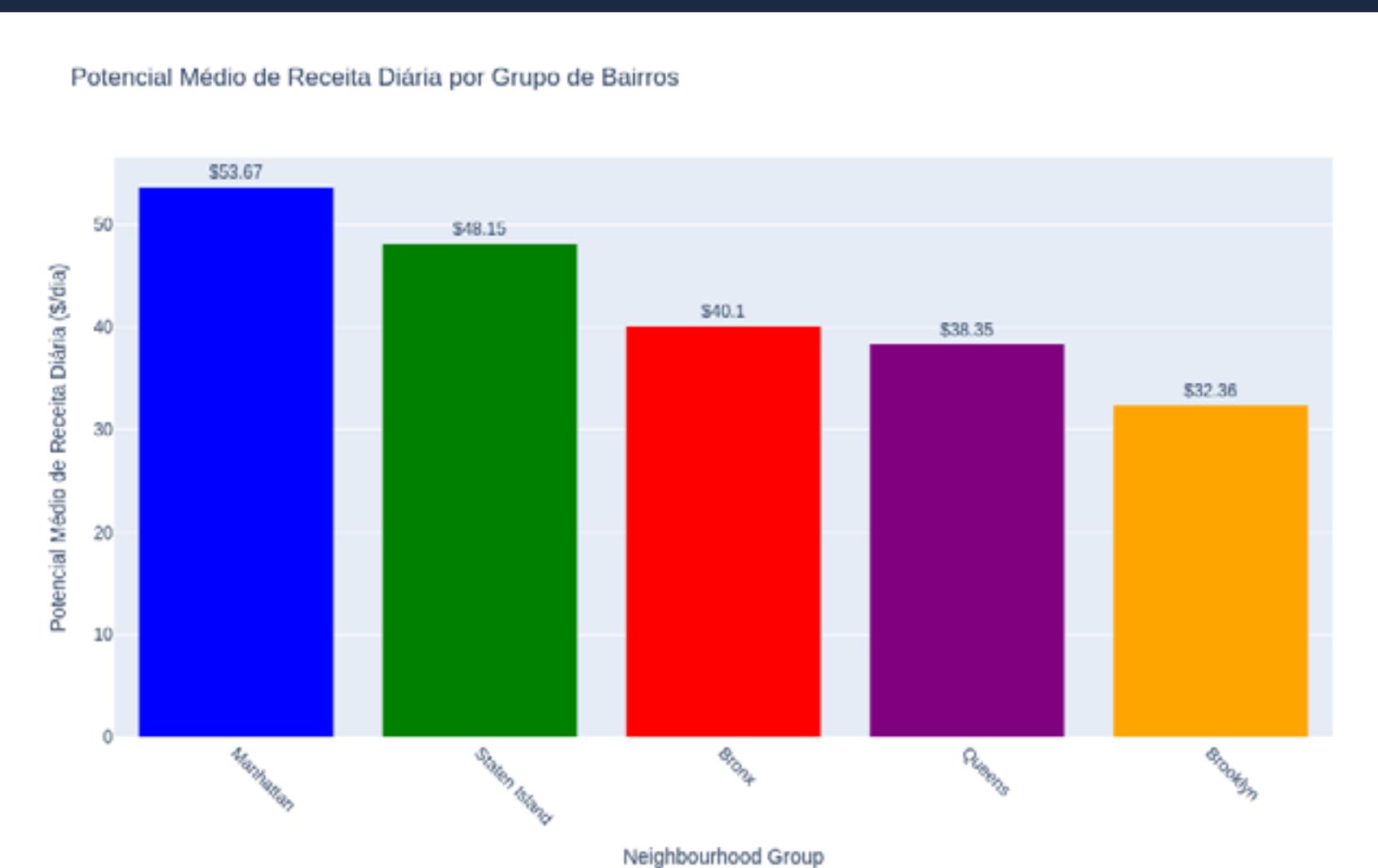
ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• GRUPO DE BAIRROS



Ao comparar as médias dos preços e da disponibilidade por região dos bairros, vemos que nas regiões onde tem menos propriedades a disponibilidade ao longo do ano é maior, porém a média de preço é menor.

Já nas duas regiões onde as propriedades apresentam uma menor disponibilidade, a média de preço supera as demais regiões.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

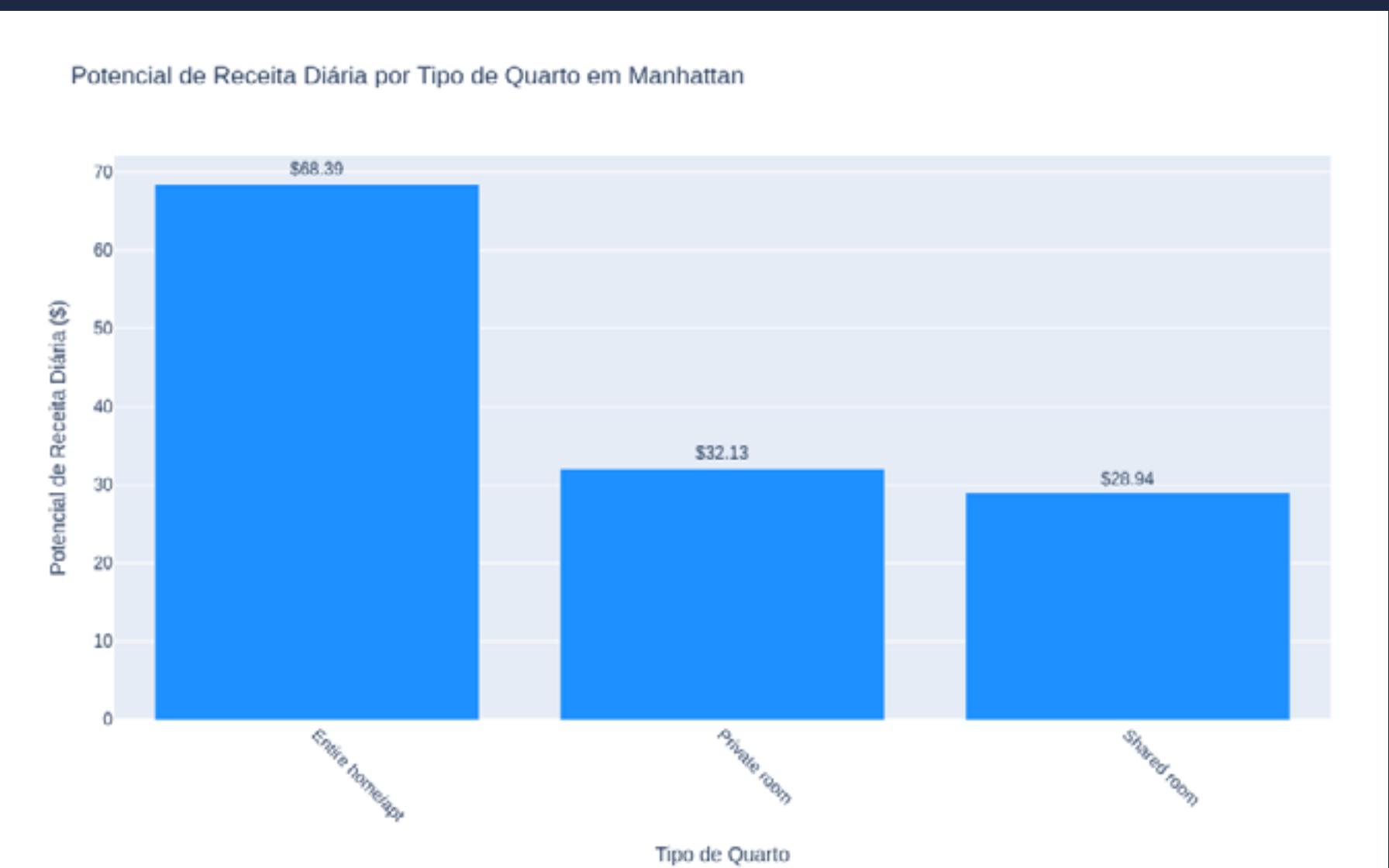
ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• MANHATTAN



Sabendo que Manhattan tem o melhor Potencial de Receita Diária, vemos no gráfico ao lado o melhor tipo de propriedade.

Comparando com o Potencial de todos os bairros, Manhattan apresenta valores levemente superiores, o que pode ser interessante para um possível investidor, já que o mesmo deve pensar se o valor da propriedade irá trazer lucros. Mas outra alternativa, de acordo com o valor das propriedades, seria este adquirir duas propriedades em outras regiões.

Esta é a resposta para o item 2 das entregas.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

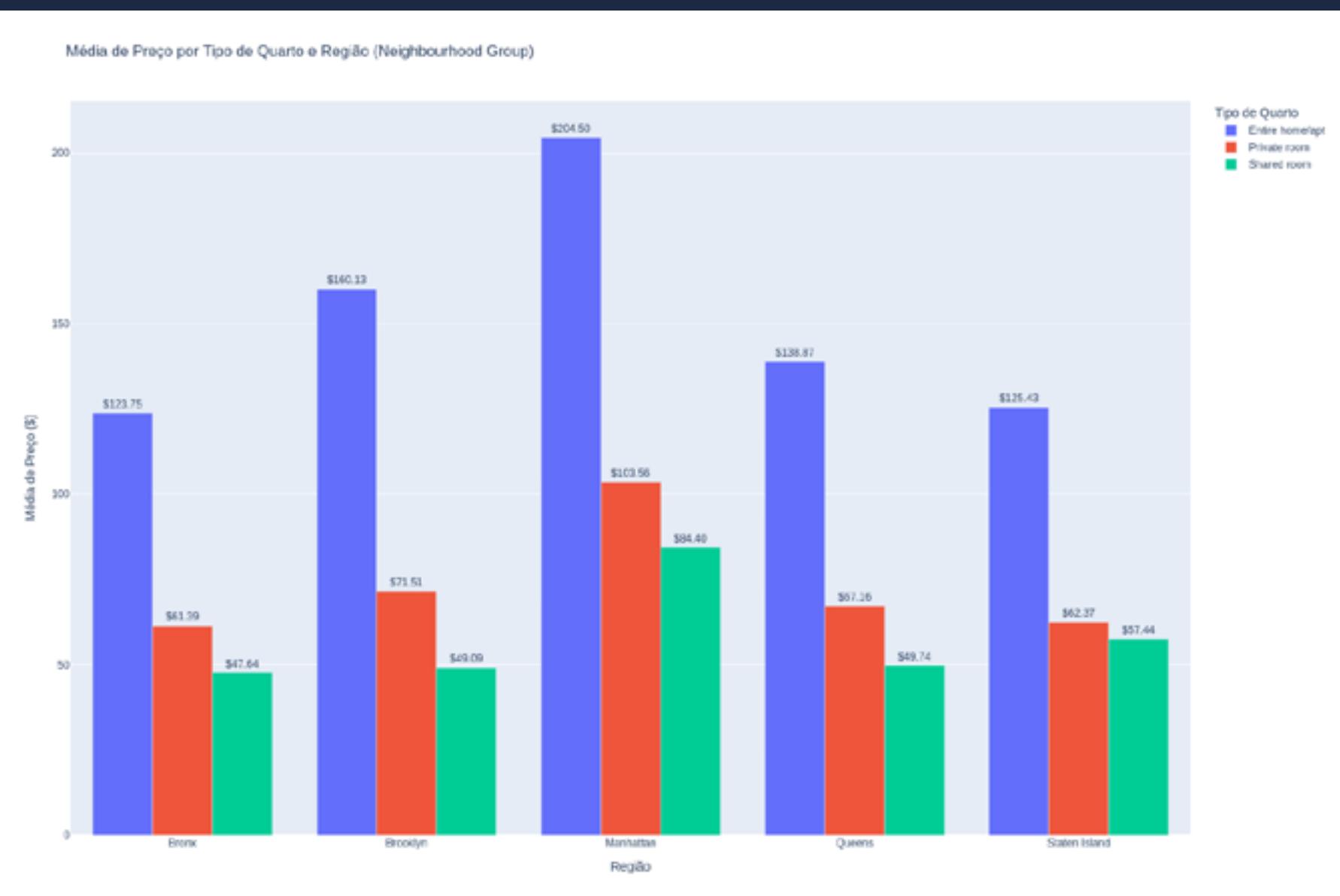
ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

- TIPOS DE QUARTO (PROPRIEDADES)



Analisando o gráfico ao lado, podemos afirmar (e respondendo a pergunta feita a pouco) que em todos os bairros, só compensa alugar o imóvel inteiro, caso o mesmo tenha 1 ou dois quartos, pois os valores ficam muito próximos.

Agora, caso tenha 3 ou mais, o melhor é alugar eles individualmente, como quarto privativo ou compartilhado, pois o proprietário teria um acréscimo na receita considerando o mesmo período.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

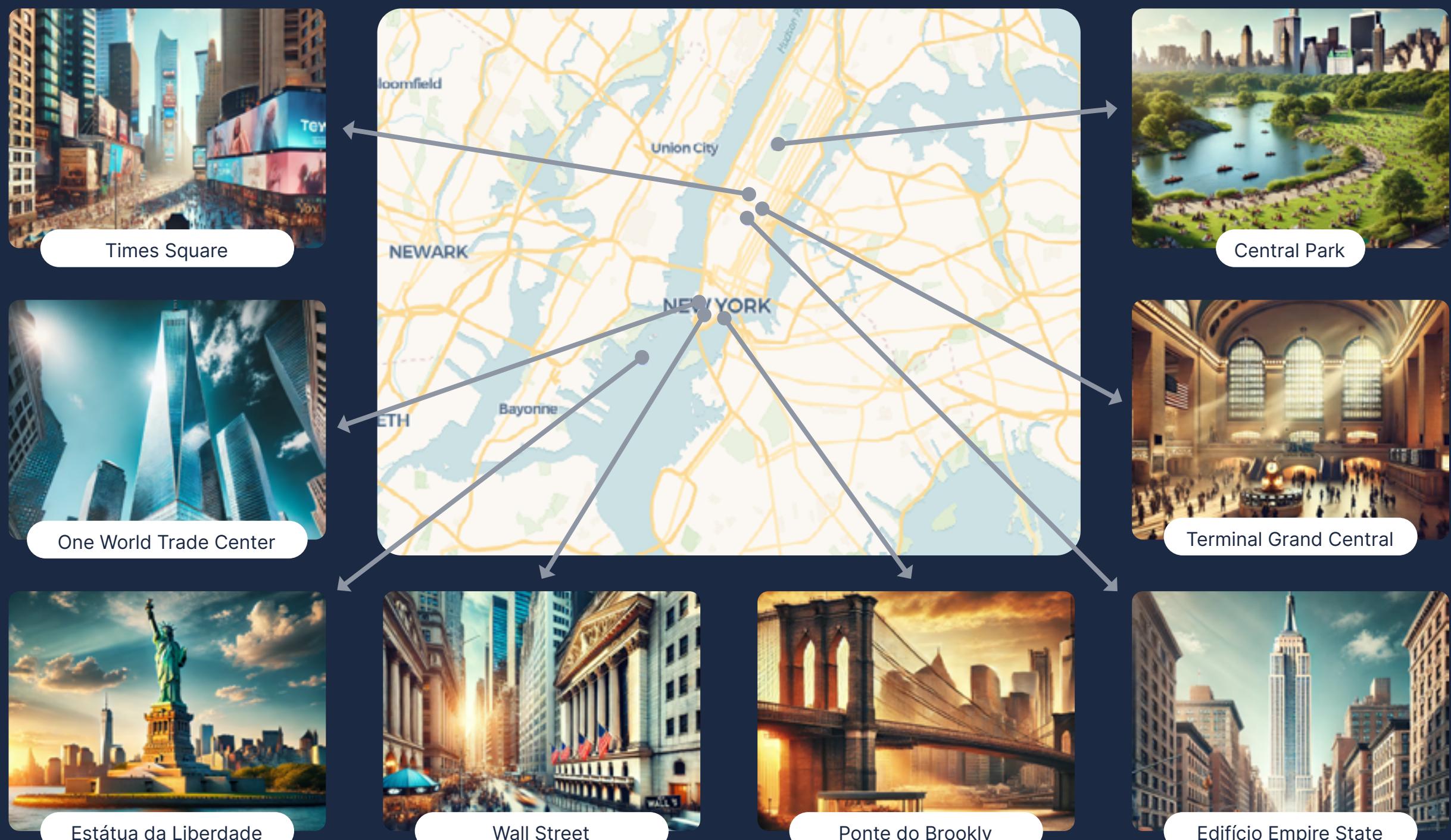
- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• MAPAS

Nestas análises separei os valores na variável price em quartis.

O objetivo é identificar com gráficos de dispersão se o preço tem alguma relação com algumas atrações turísticas, sendo elas:



DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

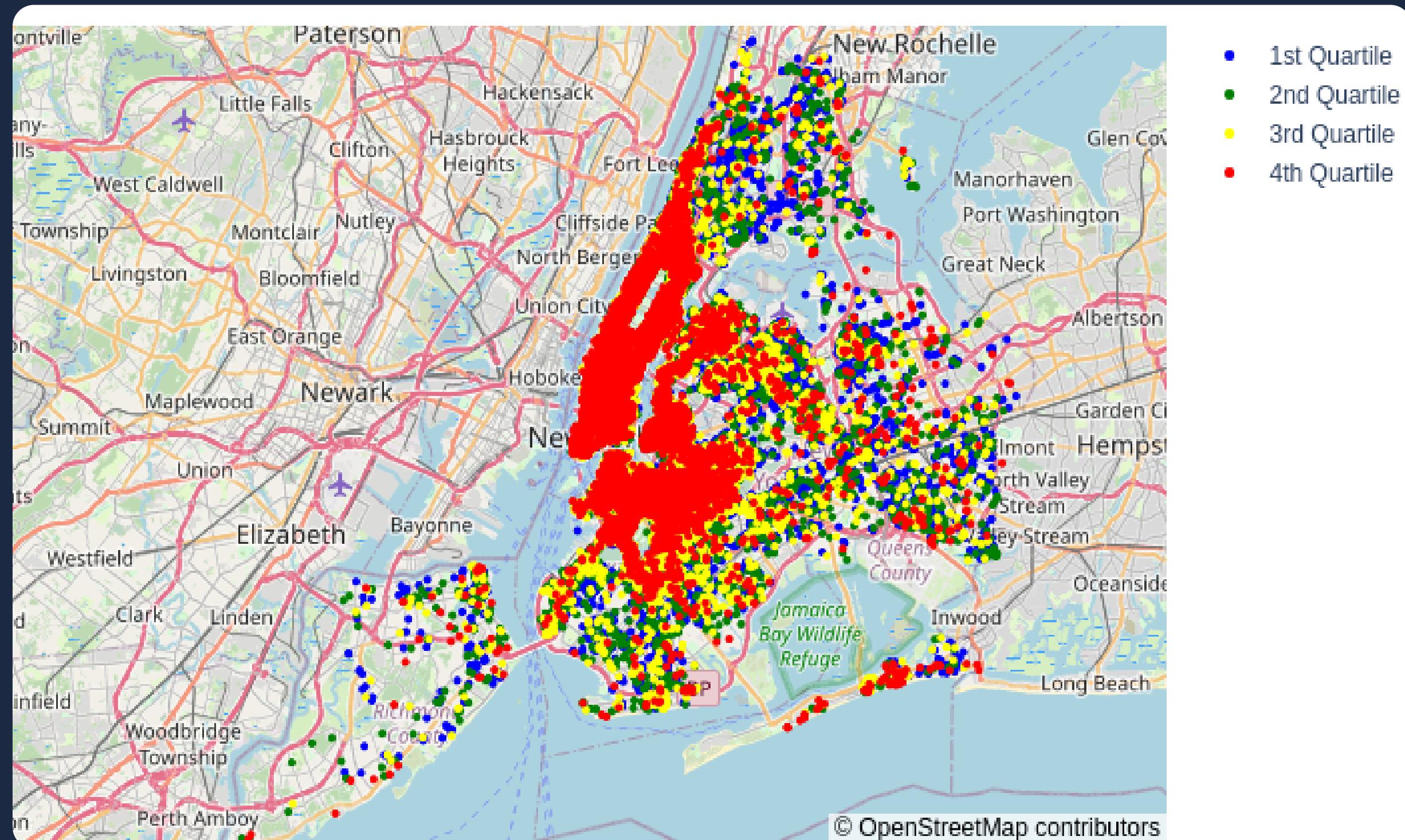
- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• MAPAS

Esta imagem mostra todas as propriedades diferenciadas por cor.

Já podemos ver como o 4º quartil (com os preços mais elevados) se concentram próximo a região central. Mas isso fica ainda visível quando...



DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

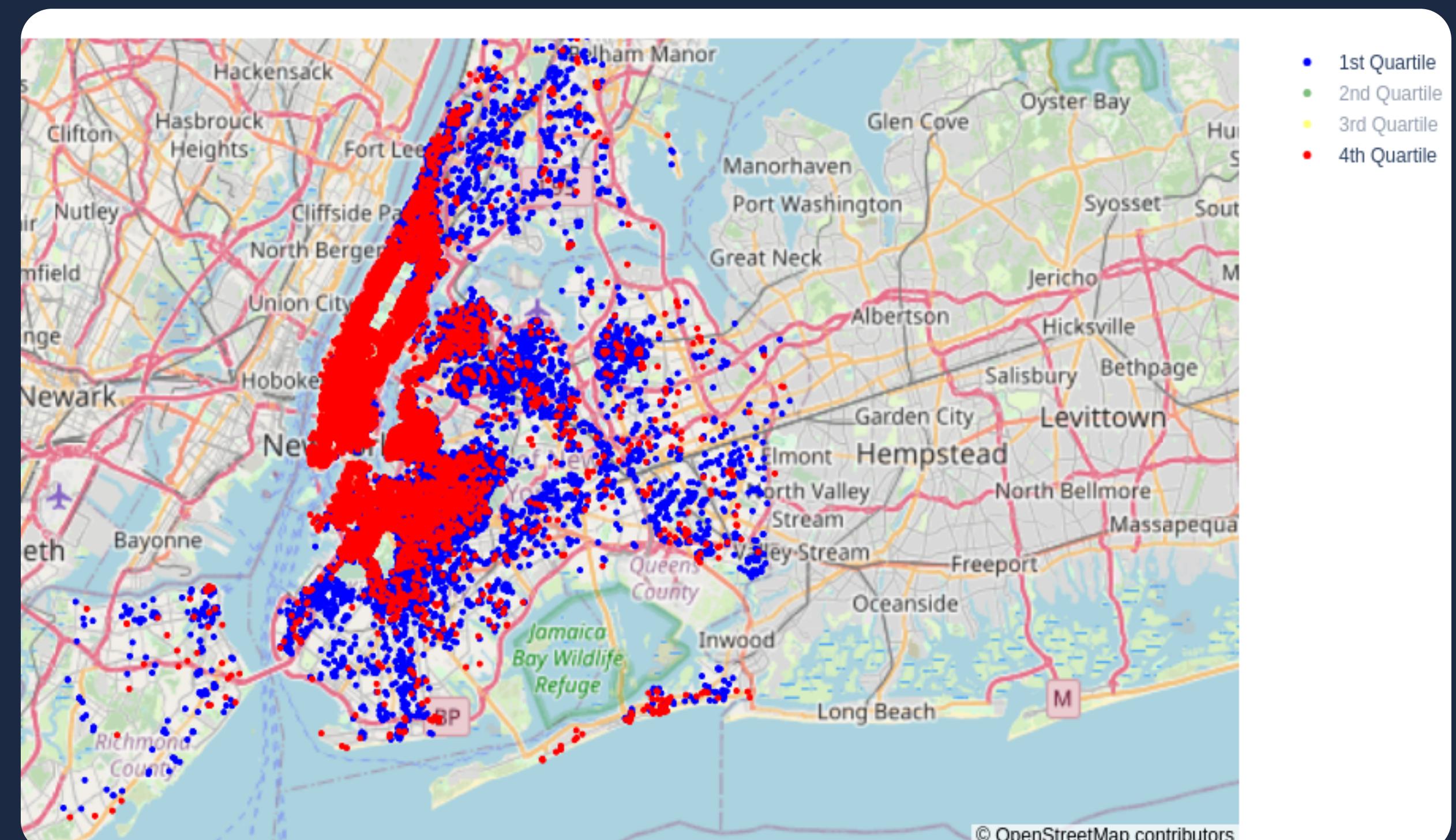
- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• MAPAS

...comparamos apenas o 1º e 4º quartil.

Exceto a região central, podemos ver poucas propriedades longe da área central e estas estão localizadas de proximas a Estátua da liberdade e em uma região litorânea.



DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

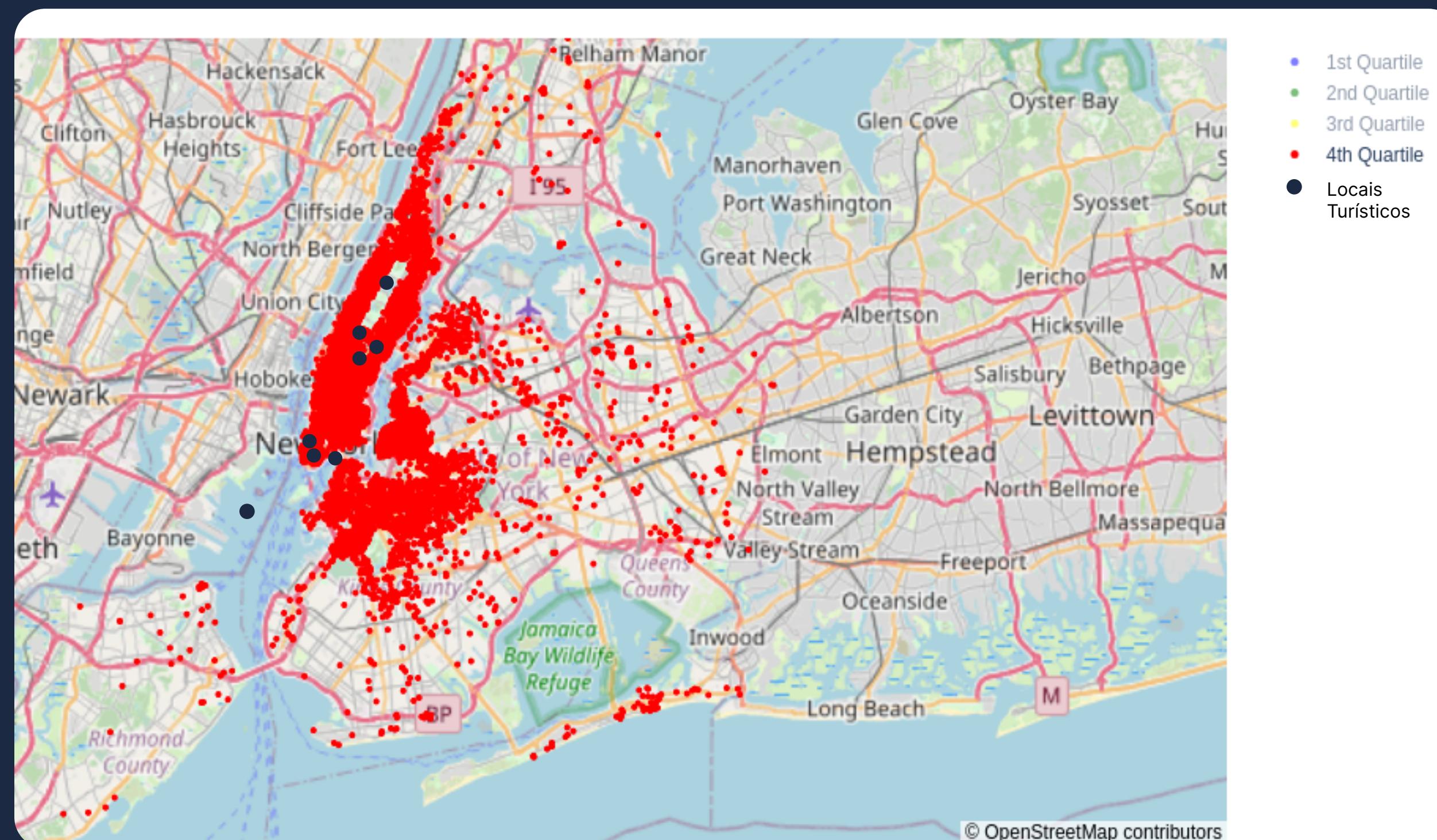
ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• MAPAS

Ao incluir os pontos turísticos, podemos considerar que os pontos turísticos tem uma relevância nos valores das propriedades.



DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

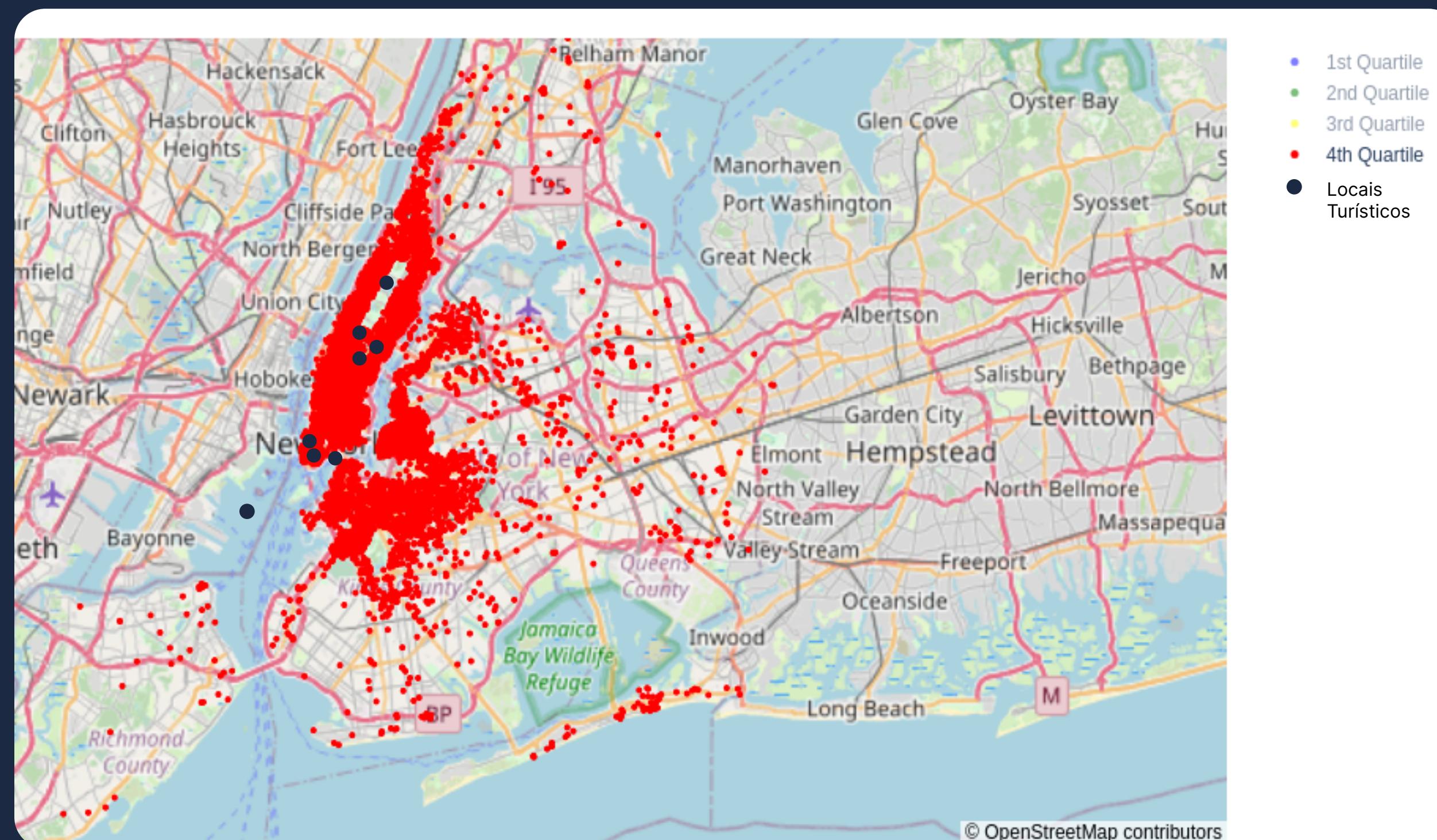
ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• MAPAS

Ao incluir os pontos turísticos, podemos considerar que os pontos turísticos tem uma relevância nos valores das propriedades.



DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• CENTRAL PARK



Para finalizar esta EDA, vou pesquisar se existe alguma ligação do nome da propriedade com o preço.

Através do mapa, foi possível visualizar que as propriedades que ficam próximas ao Central Park possuem preços mais altos, preços que ficam no 4º Quartil.

Mas será que os usuários utilizam essa expressão Central Park para atrair mais interessados?

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• CENTRAL PARK

```
# Filtrar as propriedades que estão nos quartis 1, 2 e 3
quartile_1_2_3_data = df_map[df_map['price_quartile'].isin(['1st Quartile', '2nd Quartile', '3rd Quartile'])]

# Verificar a ocorrência de "Central Park" na coluna 'name'
contains_central_park = quartile_1_2_3_data['name'].str.contains('Central Park', case=False, na=False)

# Contar quantas vezes a frase "Central Park" aparece
central_park_count = contains_central_park.sum()

# Calcular o total de propriedades nos quartis 1, 2 e 3
total_properties_in_quartiles = quartile_1_2_3_data.shape[0]

# Calcular o percentual de propriedades que mencionam "Central Park"
percent_central_park = (central_park_count / total_properties_in_quartiles) * 100
```

Na primeira análise, fiz um filtro para saber quantas vezes aparece a expressão Central Park somente no 4º Quartil do preço.

A resposta foi que a mesma aparece 460 vezes, o que representa 3,91% das propriedades no 4º quartil

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• CENTRAL PARK

```
# Filtrar as propriedades que estão nos quartis 1, 2 e 3
quartile_1_2_3_data = df_map[df_map['price_quartile'].isin(['1st Quartile', '2nd Quartile', '3rd Quartile'])]

# Verificar a ocorrência de "Central Park" na coluna 'name'
contains_central_park = quartile_1_2_3_data['name'].str.contains('Central Park', case=False, na=False)

# Contar quantas vezes a frase "Central Park" aparece
central_park_count = contains_central_park.sum()

# Calcular o total de propriedades nos quartis 1, 2 e 3
total_properties_in_quartiles = quartile_1_2_3_data.shape[0]

# Calcular o percentual de propriedades que mencionam "Central Park"
percent_central_park = (central_park_count / total_properties_in_quartiles) * 100
```

Já do 1º ao 3º quartil, a expressão 'Central Park' aparece 847 o que corresponde a 2.40% das propriedades nos quartis 1, 2 e 3.

DESAFIO

Objetivo Deste Desafio

Entregas Esperadas

Informações Gerais

ETAPAS

Etapas do Projeto

NEGÓCIO

Entendimento do Negócio

DADOS

TRATAMENTO DOS DADOS

- Boas Práticas
- Bibliotecas Utilizadas
- Dataset
- Padronização
- Dados Faltantes
- Outliers
- Correlação

ANÁLISES

ANÁLISE EXPLORATÓRIA DOS DADOS

- Grupo de Bairros (Proporção)
- Potencial Receita por Tipo de Propriedade
- Média de Preço e Disponibilidade nos Grupos de Bairros
- Grupo de Bairros
- Manhattan
- Tipos de Quarto (Propriedades)
- Mapas
- Central Park

ANÁLISE EXPLORATÓRIA DOS DADOS

• CENTRAL PARK



Conforme mostrei a pouco, a expressão Central Park esta sendo utilizada tanto para propriedades mais caras como nas demais faixas de valor.

Porém nas propriedades de maior valor a expressão é usada, estatisticamente, mais vezes.

Programa LIGHTHOUSE

indicium

Desafio Cientista de Dados - NY Rental



Luiz Augusto Soutes



48 99682.0878

in soutes
git soutes

✉️ soutes@gmail.com