

# WareBnb

Data Warehouse para análise do impacto das características do imóvel no seu preço.

Acauã Pitta Corrêa da Silva

Luiz Gustavo Abou Hatem

# Introdução

O setor de aluguel de curto prazo, exemplificado pelo AirBnb, tem crescido exponencialmente nos últimos anos. Com isso, a compreensão dos padrões de preços dos imóveis se torna fundamental tanto para os proprietários e investidores, que desejam maximizar seus lucros, quanto para os viajantes, que buscam acomodações que atendam às suas necessidades e orçamento.



# Problema de Análise

O problema específico que enfrentamos é entender quais os fatores que influenciam os preços dos imóveis listados no AirBnb. Identificar quais características dos imóveis, localizações geográficas e outras variáveis impactam de maneira significativa nos preços de aluguel.



# Público Alvo

Anfitriões:

- O que é mais lucrativo investir no imóvel para poder cobrar mais?
- É importante morar perto do imóvel?
- Vale a pena alugar o imóvel por longos períodos ou curtos nessa localização?



# Público Alvo

Hóspedes:

- Quais os bairros mais baratos?
- Vale a pena alugar um imóvel com mais banheiros?
- Quanto o hóspede precisa pagar a mais (em média) para alugar em uma região mais bem avaliada?



# Público Alvo

Investidores:

- Qual bairro gera maior receita?
- Vale mais a pena comprar imóveis em bairros litorâneos ou centrais?
- Quanto que os cômodos e os banheiros do imóvel afetam na receita?



# Coleções de Dados

## **Fontes de Dados:**

Utilizaremos os dados abertos fornecidos pelo AirBnb, que incluem informações detalhadas sobre os imóveis listados na cidade do Rio de Janeiro, coletadas em Dezembro de 2023.

## **Descrições dos Conteúdos:**

Informações sobre os imóveis, preços, localização geográfica, avaliações dos hóspedes, etc.

## **Link:**

<https://data.insideairbnb.com/brazil/rj/rio-de-janeiro/2023-12-26/data/listings.csv.gz>



# Esquema Dimensional (Estrela)

## D. Localização

id (pk)  
neighbourhood (varchar(200))  
latitude(numeric(8,5))  
longitude(numeric(8,5))

## Fato Preço

id\_localizacao (pk, fk)  
price (numeric(8,2))

## D. Propriedade

id (pk)  
property\_type(vARCHAR(100))  
room\_type(vARCHAR(100))  
accommodates(numeric(4))  
bathrooms(numeric(4,2))  
beds(numeric(4))

## D.Host

id\_dimension\_host (pk)  
host\_acceptance\_rate(numeric(4,2))  
host\_is\_superhost (boolean)  
host\_neighbourhood (varchar(100))  
host\_listings\_count (numeric(4))

## D. Reviews

id (pk)  
number\_of\_reviews(numeric(8))  
review\_scores\_rating(numeric(4,2))  
review\_scores\_accuracy(numeric(4,2))  
review\_scores\_cleanliness(numeric(4,2))  
review\_scores\_checkin(numeric(4,2))  
review\_scores\_communication(numeric(4,2))  
review\_scores\_location(numeric(4,2))  
review\_scores\_value(numeric(4,2))



# Dicionário de Atributos

## *Tabela Dimensão Localização*

**id\_dimension\_local (pk):** Identificador único da localização.

**neighbourhood:** Nome do bairro onde o imóvel está localizado.

**latitude:** Latitude do imóvel.

**longitude:** Longitude do imóvel.



# Dicionário de Atributos

## *Tabela Dimensão Host*

**id\_dimension\_host (pk):** Identificador único da data.

**host\_acceptance\_rate:** Taxa de requisições aceitas pelo usuário.

**host\_is\_superhost:** Se o host é superhost, um host recebe o mérito de superhost quando obtém muitas avaliações positivas

**host\_neighbourhood:** Onde o host mora.

**host\_listings\_count:** Quantos imóveis são listados por esse host, em outras palavras, quantos imóveis esse host atende.



# Dicionário de Atributos

## *Tabela Dimensão Propriedade*

**id\_dimension\_property (pk):** Identificador único da propriedade.

**property\_type:** Tipo de propriedade (apartamento, casa, etc.).

**room\_type:** Tipo de quarto (quarto inteiro, quarto compartilhado, etc.).

**accommodates:** Capacidade de acomodação da propriedade.

**bathrooms:** Número de banheiros na propriedade. (0.5 se for lavabo)

**beds:** Número de camas na propriedade.



# Dicionário de Atributos

## *Tabela Dimensão Review*

**id\_dimension\_review (fk):** Identificador único de Review.

**number\_of\_reviews:** Número de reviews do imóvel.

**review\_scores\_rating:** Nota geral do imóvel.

**review\_scores\_accuracy:** Nota da veracidade da descrição do imóvel.

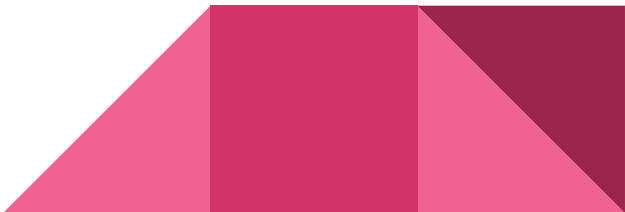
**review\_scores\_cleanliness:** Nota da limpeza do imóvel.

**review\_scores\_checkin:** Nota do check-in do imóvel, em outras palavras, da recepção e da facilidade de pegar a chave.

**review\_scores\_communication:** Nota do atendimento do anfitrião.

**review\_scores\_location:** Nota da localização do imóvel.

**review\_scores\_value:** Nota do preço do imóvel.



# Dicionário de Atributos

## *Tabela Fato Preço*

**id\_fact\_price (fk):** Chave estrangeira referenciando a dimensão Localização.

**price:** Preço médio por noite do imóvel.



# Exemplos de Consultas

- Qual a média de preços, por tipo de propriedade, em diferentes bairros?
- Qual a média de preço dos imóveis com host sendo superhost?
- Qual a média de preço dos imóveis com avaliação melhor que 4.9?
- Quais os preços médios por noite entre diferentes tipos de quartos (quarto inteiro, quarto compartilhado, etc.) em uma determinada localização?



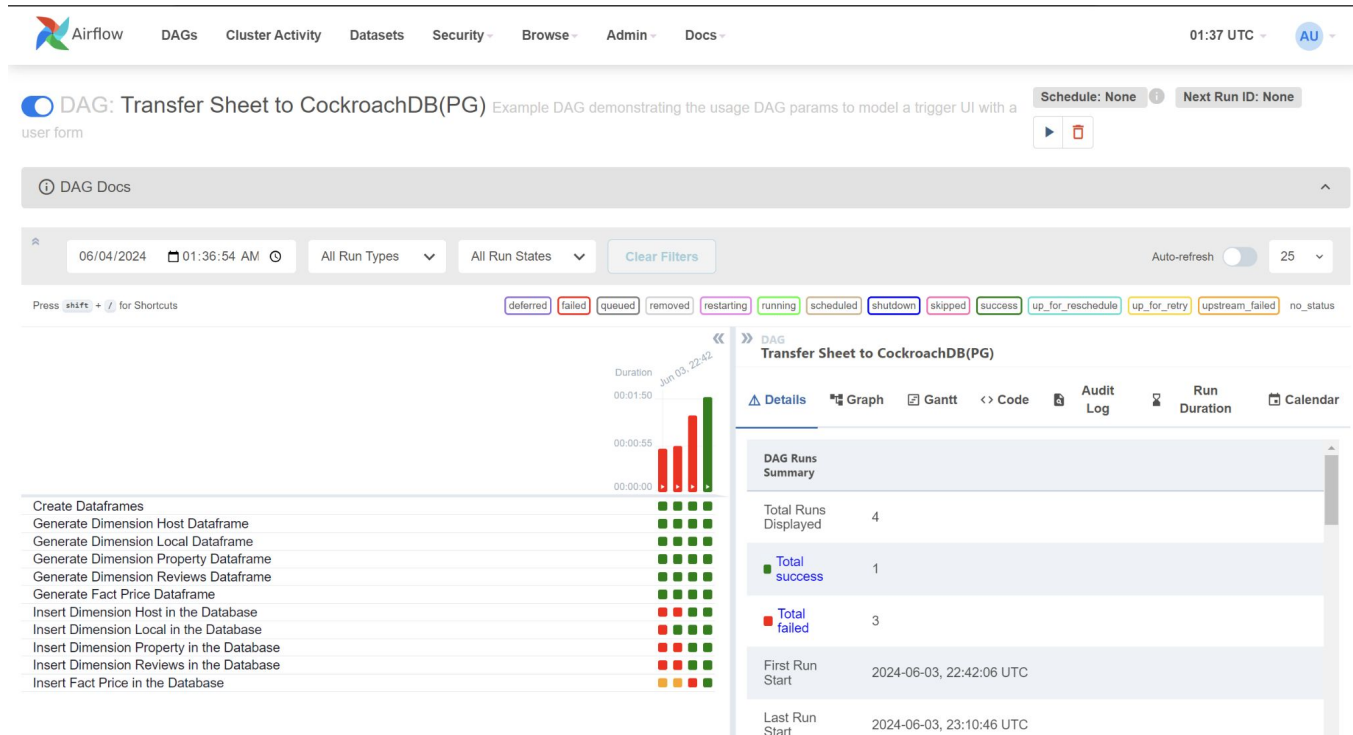
# Back-Room: Implementação

- Airflow
- Código da DAG em Python
  - Pandas para leitura de CSV e manipulação dos dados
  - SQLAlchemy para ingestão dos dados no banco
- Banco PostgreSQL\*
  - DBeaver para criação das tabelas e monitoramento dos dados
- CockroachDB (Cloud)

\* começamos com MySQL, mas migramos para PostgreSQL, pois foi o único BD que encontramos com possibilidade de ser hospedado gratuitamente sem precisar disponibilizar dados de pagamento(CockroachDB).

## Back-Room: Airflow

## Painel do Airflow






# Back-Room: Airflow

## Criação da DAG

```
with DAG(  
    dag_id=Path(__file__).stem,  
    dag_display_name="Transfer Sheet to CockroachDB(PG)",  
    schedule=None,  
    start_date=None,  
    tags=["params"],  
    params={  
        "sheet_link": Param(  
            "https://docs.google.com/spreadsheets/d/1PiSxwC-MHhwhraXVenRNvZ--ak0lvVwkYyZwmJYavY/export?format=csv",  
            type="string",  
            description="type the address to the spreadsheet with export?format=csv in the end",  
            title="Sheet Link",  
        ),  
        "pg_params": Param(  
            ["round-fairy-7625.gcp-us-east1.cockroachlabs.cloud", "26257", "defaultdb", "warebnb", "postgres", "9ArGo5tbh6WR6w-IrVfjMw"],  
            type="array",  
            description="PostgreSQL Params",  
            title="PostgreSQL Params",  
        ),  
    },  
)  
as dag:
```

# Back-Room: Airflow

## Tasks de manipulação de DataFrames (Organização e limpeza dos dados):

- create\_general\_dataframe
  - generate\_dimension\_host\_df
  - generate\_dimension\_local\_df
  - generate\_dimension\_property\_df
  - generate\_dimension\_reviews\_df
  - generate\_fact\_price\_df
- 

# Back-Room: Airflow

## Tasks de manipulação de DataFrames (Organização e limpeza dos dados):

```
@task(task_id="generate_dimension_host_df", task_display_name="Generate Dimension Host Dataframe")
def generate_dimension_host_df(**kwargs):
    ti: TaskInstance = kwargs["ti"]
    df = ti.xcom_pull(task_ids='create_general_dataframe')['df']
    df_dimension_host = df.loc[:,("id","host_acceptance_rate", "host_is_superhost", "host_neighbourhood", "host_listings_count")]
    df_dimension_host["id_dimension_host"] = df_dimension_host["id"]
    df_dimension_host['host_is_superhost'] = df_dimension_host['host_is_superhost'].map({'t': True, 'f': False})
    df_dimension_host['host_acceptance_rate'] = df_dimension_host['host_acceptance_rate'].fillna('0%')
    df_dimension_host['host_acceptance_rate'] = df_dimension_host['host_acceptance_rate'].str.replace('%', '').astype(float) / 100
    df_dimension_host['host_listings_count'] = df_dimension_host['host_listings_count'].fillna('1')
    df_dimension_host['host_listings_count'] = df_dimension_host['host_listings_count'].astype(int)
    df_dimension_host_with_correct_id_name = df_dimension_host[["id_dimension_host", "host_acceptance_rate", "host_is_superhost",
                                                                "host_neighbourhood", "host_listings_count"]]
    return {"df_dimension_host": df_dimension_host_with_correct_id_name}
```

exemplo de task de manipulação de dataframe

# Back-Room: Airflow

## Tasks de ingestão de dados:

- insert\_dimension\_host
- insert\_dimension\_local
- insert\_dimension\_property
- insert\_dimension\_reviews
- insert\_fact\_price



# Back-Room: Airflow

## Tasks de ingestão de dados:

```
@task(task_id="insert_dimension_host", task_display_name="Insert Dimension Host in the Database")
def insert_dimension_host(**kwargs):
    ti: TaskInstance = kwargs["ti"]
    dag_run: DagRun = ti.dag_run
    df_dimension_host: pd.DataFrame = ti.xcom_pull(task_ids='generate_dimension_host_df')['df_dimension_host']

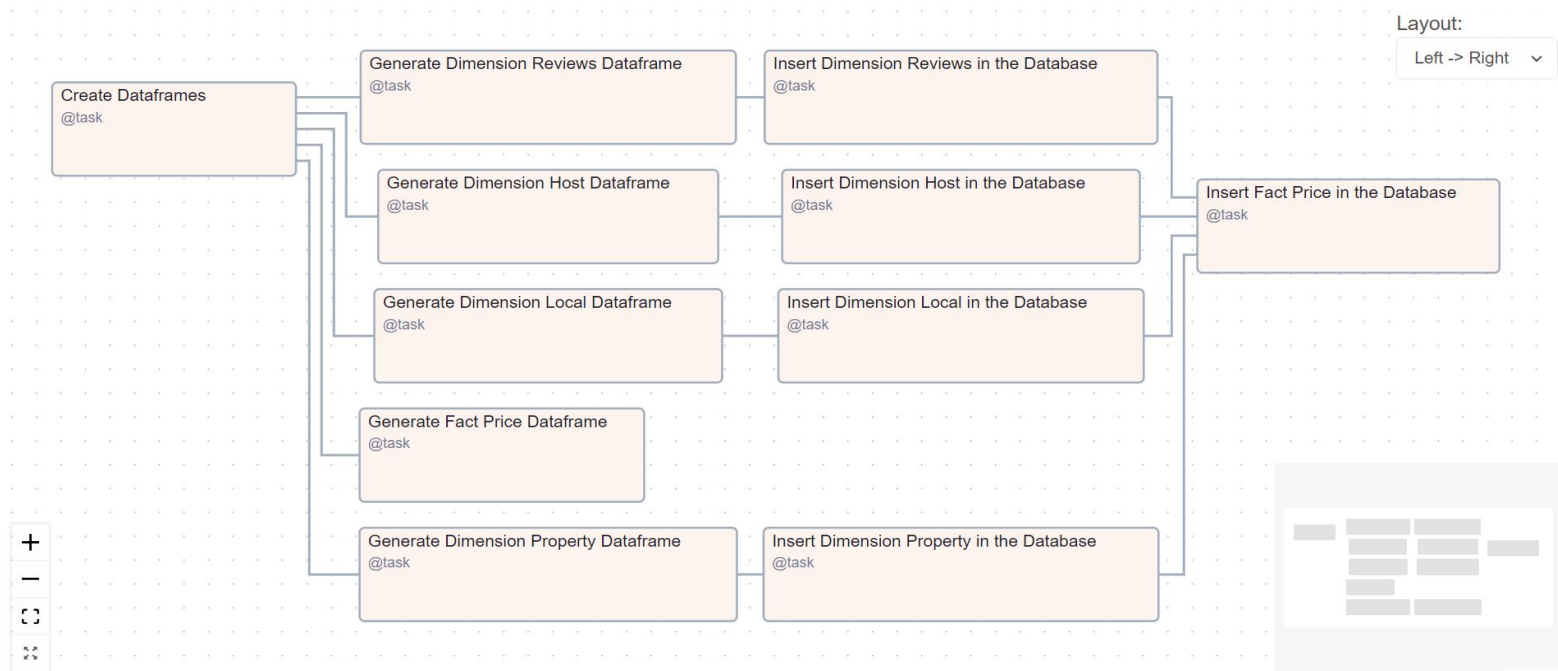
    pg_params = dag_run.conf["pg_params"]
    db_params={
        "host": pg_params[0],
        "port":pg_params[1],
        "database":pg_params[2],
        "schema": pg_params[3],
        "user": pg_params[4],
        "password":pg_params[5]
    }

    engine = sqlalchemy.create_engine(f'cockroachdb://{db_params["user"]}:{db_params["password"]}'+
                                      f'@{db_params["host"]}:{db_params["port"]}/{db_params["database"]}')
    df_dimension_host.to_sql(name='dimension_host', con=engine, if_exists='append', index=False, schema=db_params["schema"])
    return {"code": 200}
```

exemplo de task de ingestão de dados

# Back-Room: Airflow

## Ordem das Tasks:



Grafo de tarefas do airflow

# Back-Room: PostgreSQL(DBeaver)

- A escolha do DBeaver como ferramenta de administração do Banco de Dados ofereceu algumas vantagens como:
  - Versatilidade (Compatível com vários bancos de dados);
  - Ferramenta 100% Gratuita;
  - Interface Intuitiva;
  - Compatibilidade com múltiplas plataformas (Linux, Windows, MacOS);



# Back-Room: Banco PostgreSQL (Esquema e Acesso)

## Acesso:

### Host:

[round-fairy-7625.g8z.gcp-us-east1.cockroachlabs.cloud](https://round-fairy-7625.g8z.gcp-us-east1.cockroachlabs.cloud)

### Port:

26257

### Database:

defaultdb

### Schema:

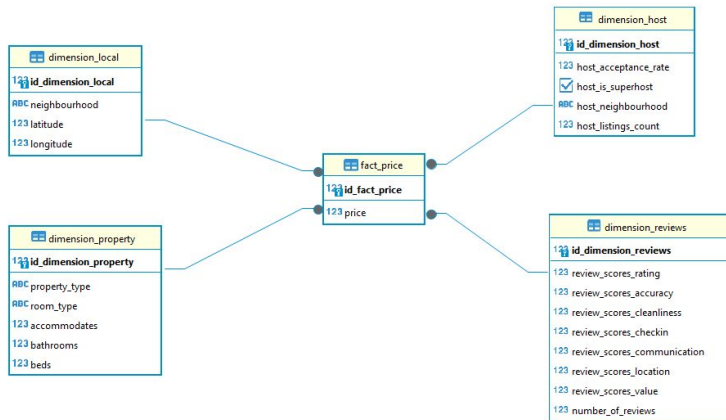
warebnb

### Username:

postgres

### Password:

9ArGo5tbh6WR6w-IrVfjMw





# FrontRoom: Consumo



## ODBC

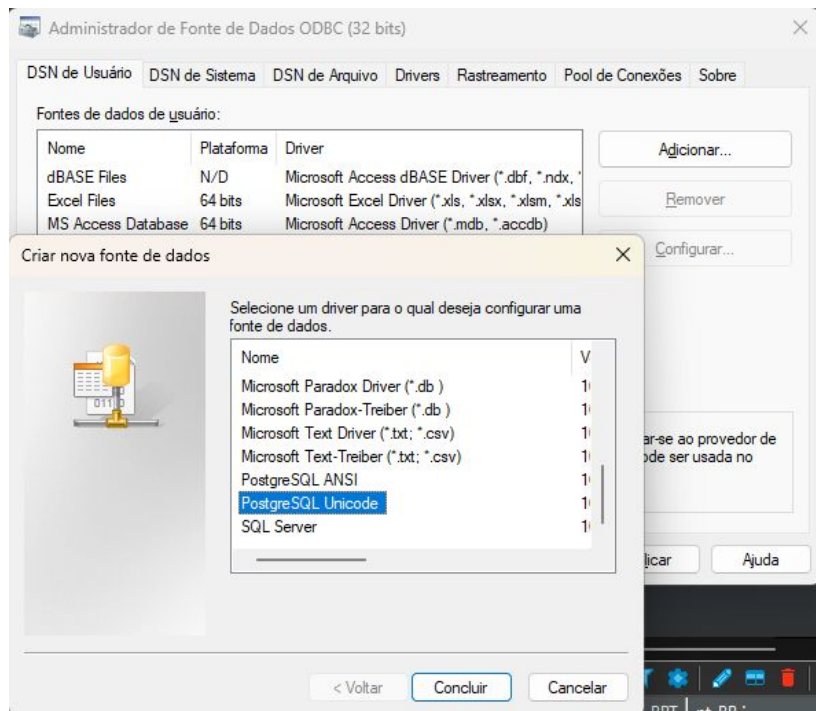
ODBC é um **protocolo** que você pode usar para conectar-se à um banco de dados. Nesse caso, utilizamos o Driver **pqslODBC** para realizar a conexão ao nosso **BD Postgres**.

## PowerBI

O Power BI se destaca como uma plataforma completa de **business intelligence (BI)** ao oferecer aos usuários uma experiência abrangente e intuitiva para gerenciar seus dados, desde a **conexão e preparação** até a **análise e visualização**, tudo em um único ambiente.

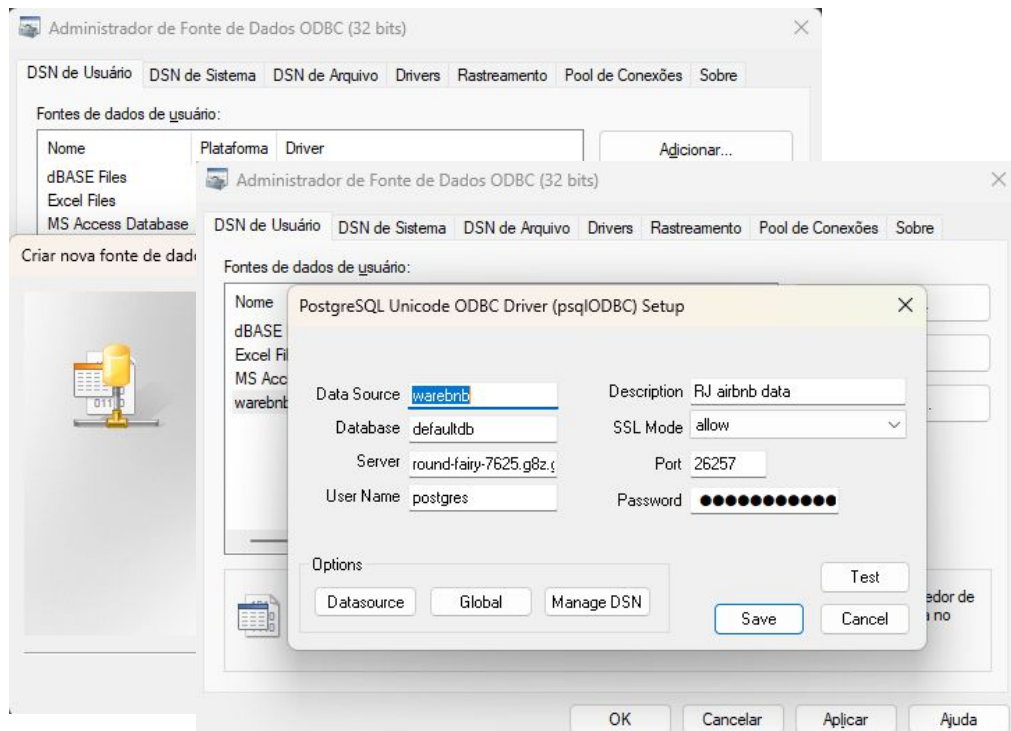


# PostgreSQL - ODBC



ODBC é um protocolo que você pode usar para conectar-se a um banco de dados. Nesse caso, utilizamos o Driver pqslODBC para realizar a conexão ao nosso BD Postgres.

# PostgreSQL - ODBC



Conexão ao DB realizada através do ODBC.

# FrontRoom: Consumo

## Importação de dados via ODBC no PowerBI

Obter Dados

ODBC

Tudo

Outro

ODBC

De ODBC

DSN (Nome da Fonte de Dados)

(Nenhum)

dBASE Files

Excel Files

MS Access Database

warebnb

(Nenhum)

OK

Cancelar

Conectores Certificados

Aplicativos de Modelo

Conectar

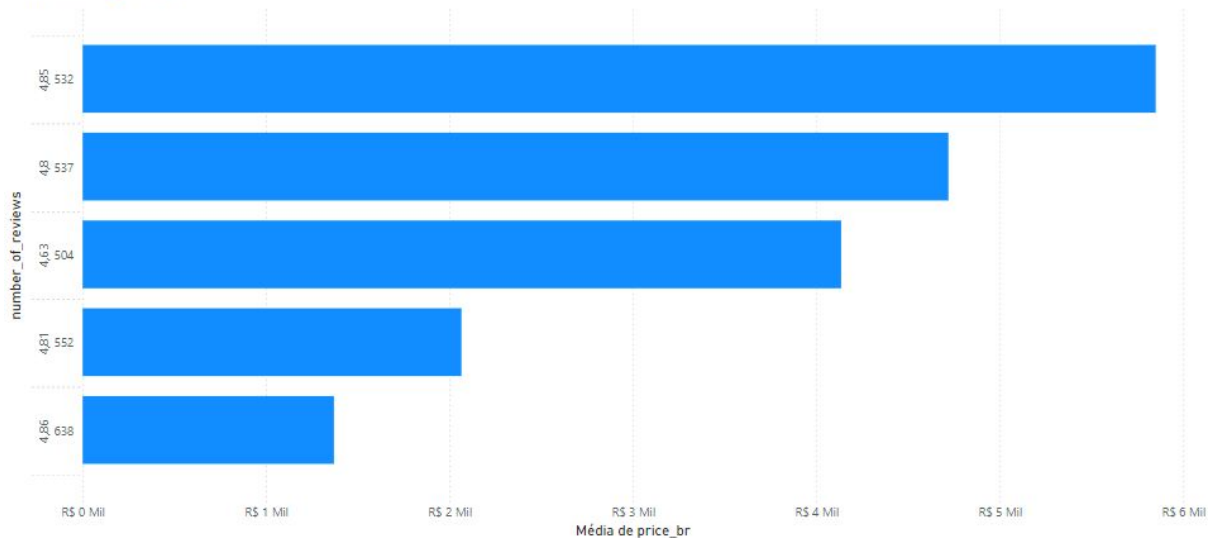
Cancelar

# Exemplos de Consultas

A consulta a seguir nos demonstra qual a **média de preço** de imóveis em **Copacabana**. Refinadas através do **número de reviews** que o imóvel recebeu, nesse caso **mais de 500**.

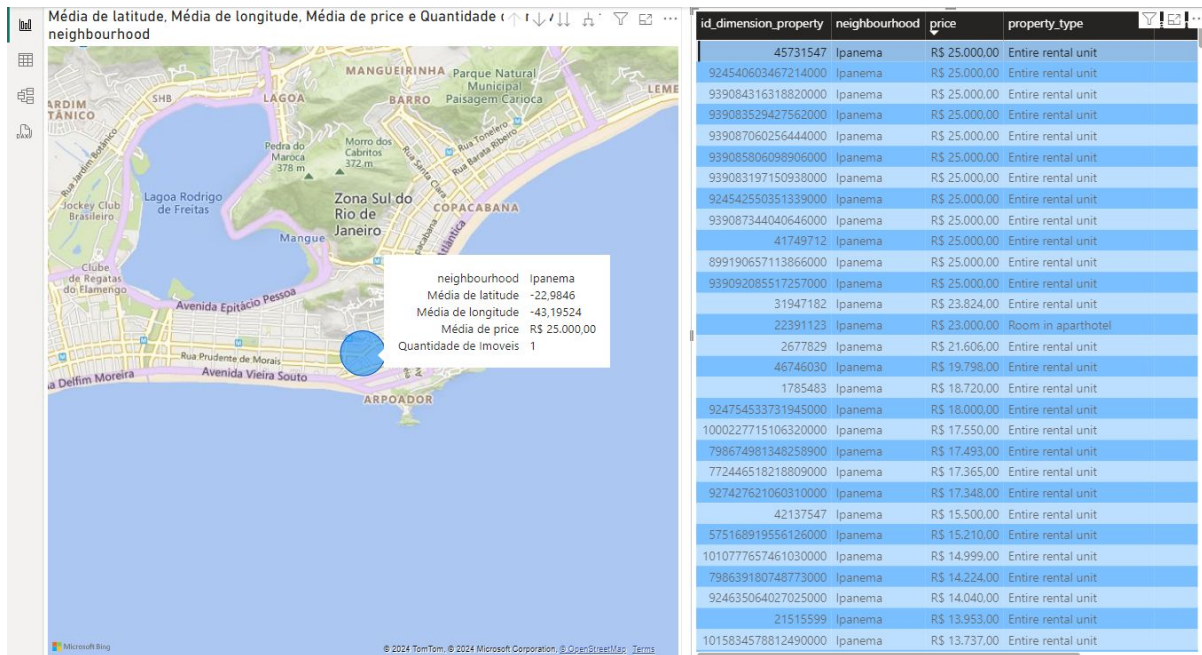
Média de price\_br por review\_scores\_rating, number\_of\_reviews e neighbourhood

neighbourhood ● Copacabana



# Exemplos de Consultas

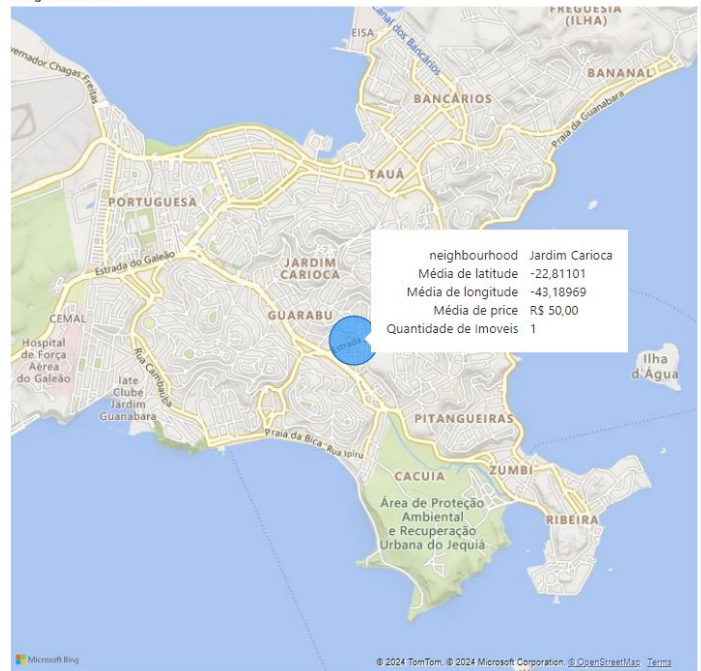
Qual o preço **máximo** da diária no bairro Ipanema?



# Exemplos de Consultas

Qual o **menor** faturamento diário por bairros?

Média de latitude, Média de longitude, Média de price e Quantidade de  
neighbourhood



property	neighbourhood	price	property_type	accommodates	ba
2184	Jardim Carioca	R\$ 50,00	Entire rental unit	2	1
0000	Padre Miguel	R\$ 50,00	Entire rental unit	4	1
2000	Vigário Geral	R\$ 51,00	Entire rental unit	2	1
9000	Campo Grande	R\$ 52,00	Entire rental unit	4	1
6610	Ramos	R\$ 55,00	Entire rental unit	2	1
1852	Santa Teresa	R\$ 56,00	Entire rental unit	2	1
2000	Campo Grande	R\$ 60,00	Entire rental unit	3	1
6000	Ribeira	R\$ 61,00	Entire rental unit	4	1
2000	Vargem Grande	R\$ 63,00	Entire rental unit	4	5
5000	Barra de Guaratiba	R\$ 64,00	Entire rental unit	2	1
6100	Praça da Bandeira	R\$ 64,00	Entire rental unit	1	1
2370	Catete	R\$ 68,00	Entire rental unit	2	1
2900	Campo Grande	R\$ 70,00	Entire rental unit	5	1
7188	Copacabana	R\$ 71,00	Entire rental unit	2	1
0000	Maré	R\$ 72,00	Entire rental unit	2	1
9000	Barra de Guaratiba	R\$ 74,00	Entire rental unit	3	1
8100	Centro	R\$ 74,00	Entire rental unit	2	1
5692	Senador Camará	R\$ 75,00	Entire rental unit	3	1
3609	Vila Isabel	R\$ 75,00	Entire rental unit	10	2
0000	Barra de Guaratiba	R\$ 79,00	Entire rental unit	6	1
3026	Centro	R\$ 80,00	Entire rental unit	4	1
4000	Recreio dos Bandeirantes	R\$ 80,00	Entire rental unit	4	1
2641	Rocinha	R\$ 80,00	Entire rental unit	1	1
7333	Santa Teresa	R\$ 80,00	Entire rental unit	3	1
8797	Vidigal	R\$ 80,00	Entire rental unit	2	1
4177	Benfica	R\$ 81,00	Entire rental unit	2	1
5116	Jardim Sulacap	R\$ 81,00	Entire rental unit	5	1
5000	Recreio dos Bandeirantes	R\$ 83,00	Entire rental unit	3	2
7769	Tijuca	R\$ 84,00	Entire rental unit	7	1

# Exemplos de Consultas

Quanto o hóspede precisa pagar **em média a mais** para alugar em uma **região mais bem avaliada**?

neighbourhood	Média de price	Média de Reviews	
Anil	R\$ 300,67	4,67	9
Méier	R\$ 421,00	4,64	7
Bonsucesso	R\$ 215,50	4,63	2
Freguesia (Ilha)	R\$ 404,50	4,63	4
Madureira	R\$ 184,50	4,63	2
Encantado	R\$ 288,00	4,60	5
Vargem Pequena	R\$ 389,37	4,58	19
Praça da Bandeira	R\$ 449,18	4,52	22
Cacuaia	R\$ 330,00	4,50	1
Catumbi	R\$ 346,00	4,50	1
Guadalupe	R\$ 221,00	4,50	1
Pedra de Guaratiba	R\$ 150,00	4,50	1
Santo Cristo	R\$ 309,33	4,50	6
Curicica	R\$ 308,88	4,44	17
Itanhangá	R\$ 430,17	4,42	6
Coelho Neto	R\$ 175,50	4,38	2
Cosme Velho	R\$ 555,00	4,36	14
Bangu	R\$ 263,00	4,33	3
Vargem Grande	R\$ 360,13	4,19	8
607901774661565100	R\$ 370,00	5,00	1
835584181660745000	R\$ 497,00	5,00	1
904568157984465000	R\$ 522,00	5,00	1
1041249053960160000	R\$ 240,00	5,00	1
706774670042210000	R\$ 208,00	4,75	1
1003013197600910000	R\$ 150,00	4,75	1
986281783700041000	R\$ 394,00	3,00	1
53328298	R\$ 500,00	1,00	1
Panque Anchieta	R\$ 300,00	4,00	1
Total	R\$ 1.137,38	4,79	15597

id_dimension_local	accommodates	beds	bathrooms	number_of_reviews	review_scores_rating	price
53328298	16	39	4	1	1,00	R\$ 500,00
607901774661565100	2	1	1	2	5,00	R\$ 370,00
706774670042210000	2	1	1	15	4,87	R\$ 208,00
835584181660745000	4	2	1	5	5,00	R\$ 497,00
904568157984465000	10	9	2	3	5,00	R\$ 522,00
986281783700041000	3	1	1	1	3,00	R\$ 394,00
1003013197600910000	4	1	1	4	4,75	R\$ 150,00
1041249053960160000	4	2	1	1	5,00	R\$ 240,00



## Vargem Grande

- média 4.19 estrelas
- 8 imóveis
- Diária média de R\$ 360,13




# Exemplos de Consultas

Quanto o hóspede precisa pagar **em média a mais** para alugar em uma **região mais bem avaliada**?

neighbourhood	Média de price	Média de Rounded	
Praça Seca	R\$ 322,67	5,00	3
Ribeira	R\$ 61,00	5,00	1
Rocinha	R\$ 286,00	5,00	1
Senador Vasconcelos	R\$ 177,00	5,00	2
Tauá	R\$ 280,50	5,00	2
Turiacú	R\$ 350,00	5,00	1
Vasco da Gama	R\$ 570,00	5,00	2
Vigário Geral	R\$ 870,50	5,00	2
Vila Valqueire	R\$ 271,00	5,00	1
Joá	R\$ 1.271,14	4,98	14
Cidade Nova	R\$ 396,00	4,95	5
Andaraí	R\$ 494,89	4,94	9
13549789	R\$ 349,00	5,00	1
45486972	R\$ 290,00	5,00	1
53640153	R\$ 742,00	5,00	1
706312875908094000	R\$ 393,00	5,00	1
882125066534725000	R\$ 223,00	5,00	1
1008647792331220000	R\$ 535,00	5,00	1
1022652843267300000	R\$ 332,00	5,00	1
44408578	R\$ 390,00	4,75	1
45487552	R\$ 1.200,00	4,75	1
Portuguesa	R\$ 156,50	4,94	4
Santa Cruz	R\$ 165,00	4,94	4
São Francisco Xavier	R\$ 285,00	4,94	4
Todos os Santos	R\$ 539,75	4,94	4
Gardênia Azul	R\$ 382,73	4,93	11
Jardim Guanabara	R\$ 405,46	4,92	13
Barra de Guaratiba	R\$ 376,05	4,92	19
Total	R\$ 1.137,38	4,79	15597

id_dimension_local	accommodates	beds	bathrooms	number_of_reviews	review_scores_rating	price
13549789	3	1	1	106	5,00	R\$ 349,00
44408578	3	2	1	41	4,73	R\$ 390,00
45486972	4	3	1	1	5,00	R\$ 290,00
45487552	5	4	1	28	4,71	R\$ 1.200,00
53640153	4	2	2	30	5,00	R\$ 742,00
706312875908094000	2	1	2	21	5,00	R\$ 393,00
882125066534725000	4	2	2	41	5,00	R\$ 223,00
1008647792331220000	4	2	1	1	5,00	R\$ 535,00
1022652843267300000	4	2	1	1	5,00	R\$ 332,00

latitude e longitude



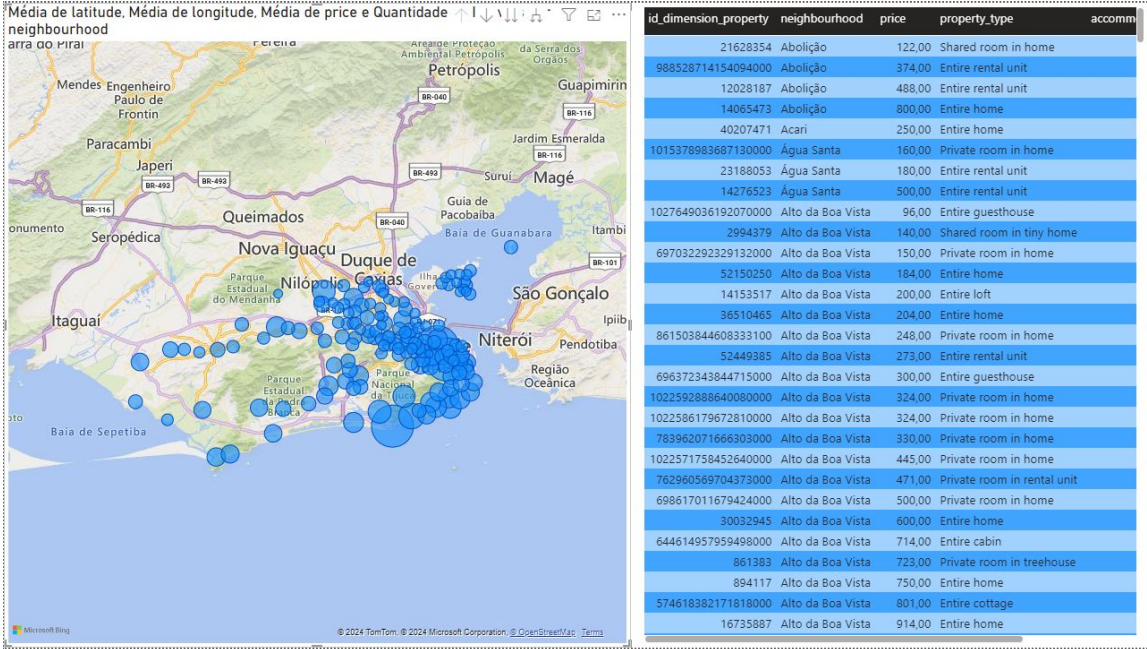
## Andaraí

- Média 4.94 estrelas
- 9 Imóveis
- Diária média de R\$ 494,89

Aproximadamente 37% mais caro

# Exemplos de Consultas

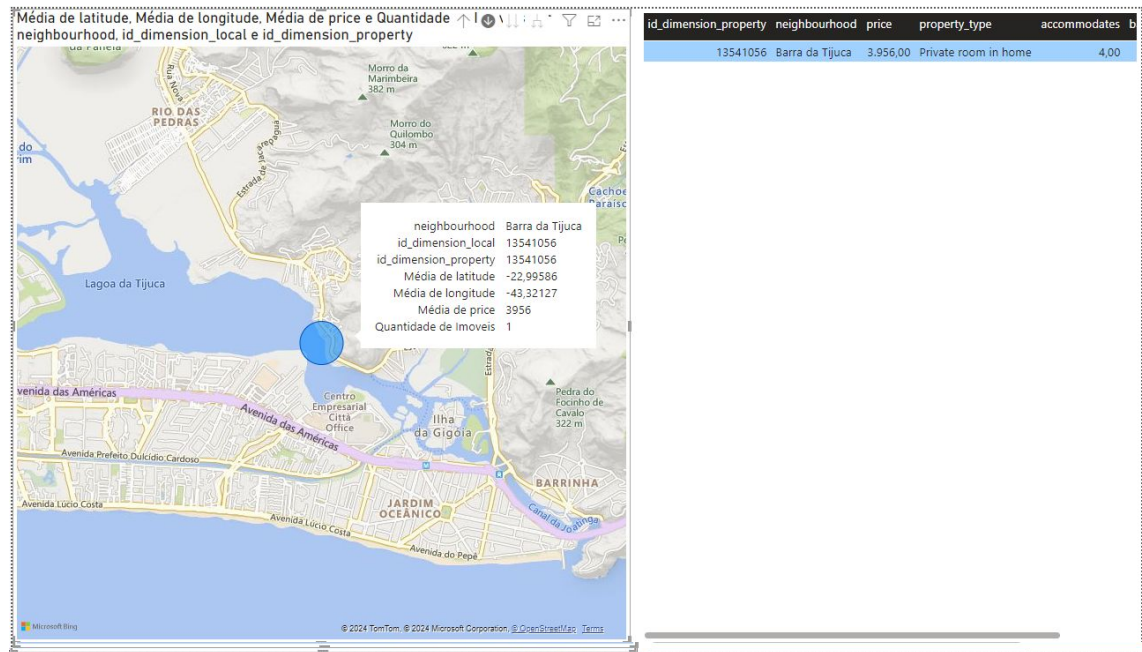
A visualização a seguir nos permite realizar **drill-down/up** sobre o **imóveis do RJ** onde a **bolha** é a **média de preço**.





# Exemplos de Consultas

E por fim, o **imóvel específico**, com as respectivas propriedades refletidas na tabela ao lado.



# Trabalhos Futuros

- Adicionar imóveis de mais cidades no Banco de Dados
- Usar Inteligência Artificial para sugerir o valor de um imóvel a ser anunciado.



# FIM

