



SOLUTION SPRINT

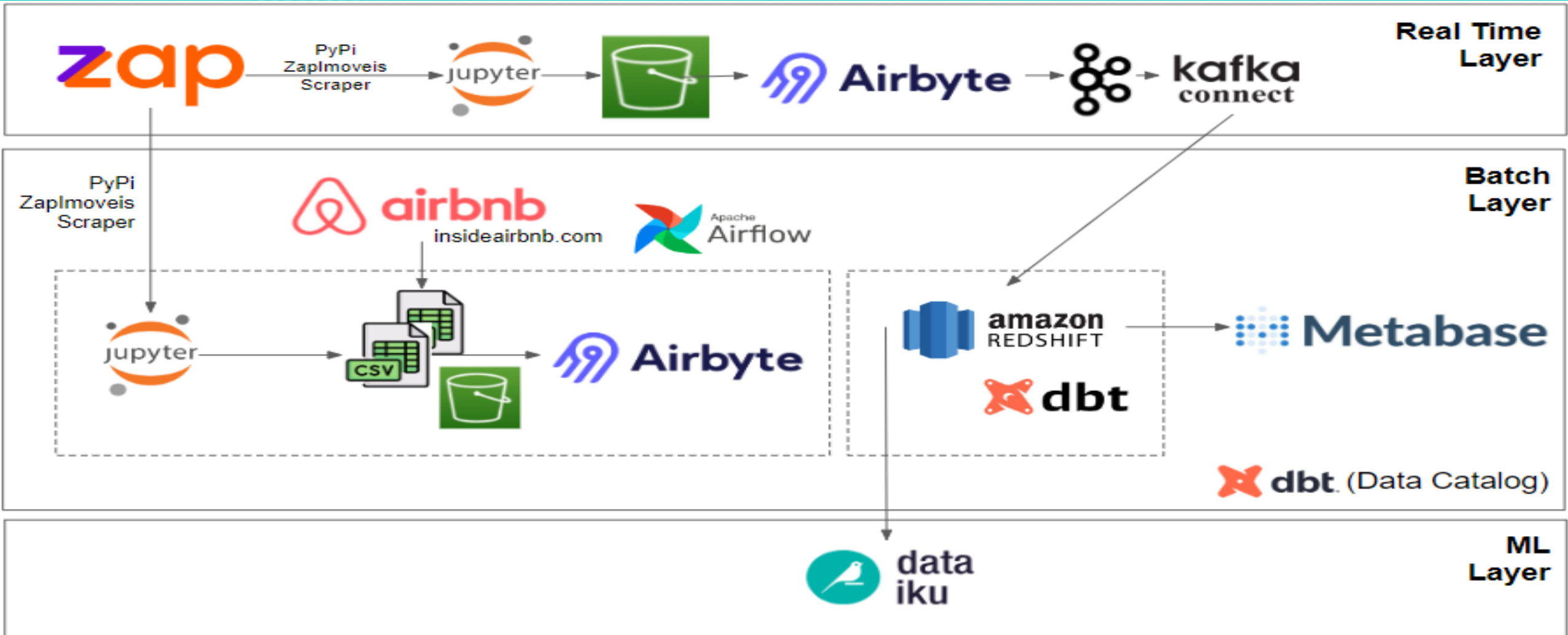
MBA ENGENHARIA DE DADOS – FASE 4

Equipe:
Marcelo Canabrava
Marina Coutinho

Descrição completa do projeto no github

[https://github.com/mcanabrava/
airbnb-zap-scrapping-ml-mba-fiap](https://github.com/mcanabrava/airbnb-zap-scrapping-ml-mba-fiap)

Diagrama da Solução



Armazenamento com upload via boto3 para o S3



zap_bs4.csv

fix scrapper area

zap_imoveis_rj.csv

add scrapper code



calendar.zip

dataset files

listings.csv

etl completed

listings_sample.csv

dataset files

reviews_sample.csv

dataset files

```
In [12]: ## CREATING THE BUCKETS

bucket_names = ['airbnb-data-landing-fiap', 'zap-data-landing-fiap']

for bucket_name in bucket_names:
    try:
        s3.create_bucket(Bucket=bucket_name)
        print(f"Bucket '{bucket_name}' created successfully.")
    except Exception as e:
        print(f"Error creating bucket '{bucket_name}': {str(e)}")
```

Bucket 'airbnb-data-landing-fiap' created successfully.
Bucket 'zap-data-landing-fiap' created successfully.


```
In [17]: ## UPLOADING ZAP DATA TO THE BUCKET

bucket_name = 'zap-data-landing-fiap'
file_path = 'dataset/zap/zap_bs4.csv'
s3_file_name = 'zap_bs4.csv'


if s3.Bucket(bucket_name) in s3.buckets.all():
    bucket = s3.Bucket(bucket_name)
    existing_objects = list(bucket.objects.filter(Prefix='zap/'))
    if any(obj['Key'] == s3_file_name for obj in existing_objects):
        print(f"A file with the name '{s3_file_name}' already exists in the '{bucket_name}' bucket.")
    else:
        try:
            s3.meta.client.upload_file(file_path, bucket_name, s3_file_name)
            print(f"File '{s3_file_name}' uploaded to '{bucket_name}' bucket successfully.")
        except Exception as e:
            print(f"Error uploading file '{s3_file_name}' to '{bucket_name}' bucket: {str(e)}")
```


File 'zap_bs4.csv' uploaded to 'zap-data-landing-fiap' bucket successfully.


Airbyte para transferência entre S3 e Redshift





FIAP PROJECT



Connections


Sources


Destinations


Builder



Billing





Resources

Your 14-day trial of Airbyte will start once your first sync has completed.

Connections / S3_RS_ZAP

S3_RS_ZAP

Enabled 

 S3  →  Redshift BETA


Status

Job History

Replication


Transformation

Settings







On time

Reset your data

 Sync now

Enabled streams

 Search

Status	Stream name	Last record loaded 	
 On time	s3_to_rs_zap	3 minutes ago	

Conferido a disponibilidade de dados no Redshift

The screenshot displays the Amazon Redshift console interface. On the left, the 'Resources Info' sidebar shows the database 'dwhdatabase' and schema 'public' selected. The main query editor, titled 'Query 1', contains the following SQL code:

```
1 SELECT
2 *
3 FROM s3_to_rs_zap
4 LIMIT 10;
```

Below the query editor, the 'Run' button is highlighted. The 'Query results' tab is active, showing the results of the query. The results are displayed in a table with the following columns:

area	city	link	garage	description
_ab_source_file_last_modified	bathrooms	_ab_source_file_url		

The table shows the first row of results, which includes the columns 'area', 'city', 'link', 'garage', and 'description'. The 'area' column contains the value '_ab_source_file_last_modified', the 'city' column contains 'bathrooms', and the 'link' column contains '_ab_source_file_url'. The 'garage' and 'description' columns are empty.

At the bottom of the console, the status of the query is shown as 'Completed, started on August 09, 2023 at 20:38:10' with an elapsed time of '00 m 02 s'. The 'Query 654' link is also visible.

Conferido a disponibilidade de dados no Redshift

Rows returned (10)

Export ▼

 Search rows

< 1 > ⚙

area ▼	city ▼	link ▼	garage ▼	description ▼	_ab_source_file_last_modifie
--------	--------	--------	----------	---------------	------------------------------

97	Rio de Janeiro	https://www.zapimoveis.com.br/imovel/venda-apartamento-2-quartos-pechincha-zona-oeste-rio-de-janeiro-rj-97m2-id-2641416694/	1	Oportunidade para realizar um bom negócio e comprar o seu apartamento!Localização privilegiada em Jacarepaguá, Estrada do Capenha próximo a Estrada do Pau Ferro.Apartamento de 97m ² com 2quartos (1 suíte), sala ampla, cozinha, área de serviço e dependência, piso em porcelanato, requinte e conforto ao seu alcance. Documentação OK!Estudamos propostas. Visitas agendadas	2023-08-09 22:48:16+00
----	----------------	---	---	--	------------------------

Modelagem de dados via DBT

The screenshot displays the DBT web interface. At the top, the DBT logo is on the left, and navigation links for 'Develop', 'Deploy', and 'Documentation' are in the center. On the right, the current project 'FiapFase4' and workspace 'Analytics' are shown, along with help and settings icons.

On the left sidebar, there are three main sections: 'test-branch' with a 'Change branch' link, 'Version Control' with a 'Create a pull request on ...' button, and 'File Explorer' showing a tree view of the project files. The 'zap_data.sql' file is highlighted in the file explorer.

The main editor area shows the 'zap_data.sql' file open. The breadcrumb navigation indicates the path: 'models > my_redshift > staging > zap_data.sql'. The SQL code in the editor is as follows:

```
1  -- dbt model: zap_data.sql
2  select *
3  from {{ source('my_redshift', 's3_to_rs_zap') }} as zap_data;
4
```

Below the editor, there is a toolbar with buttons for 'Preview', 'Compile', 'Build', and 'Format'. To the right of these are tabs for 'Results', 'Code Quality', 'Compiled Code', and 'Lineage'. The 'Lineage' tab is currently selected.

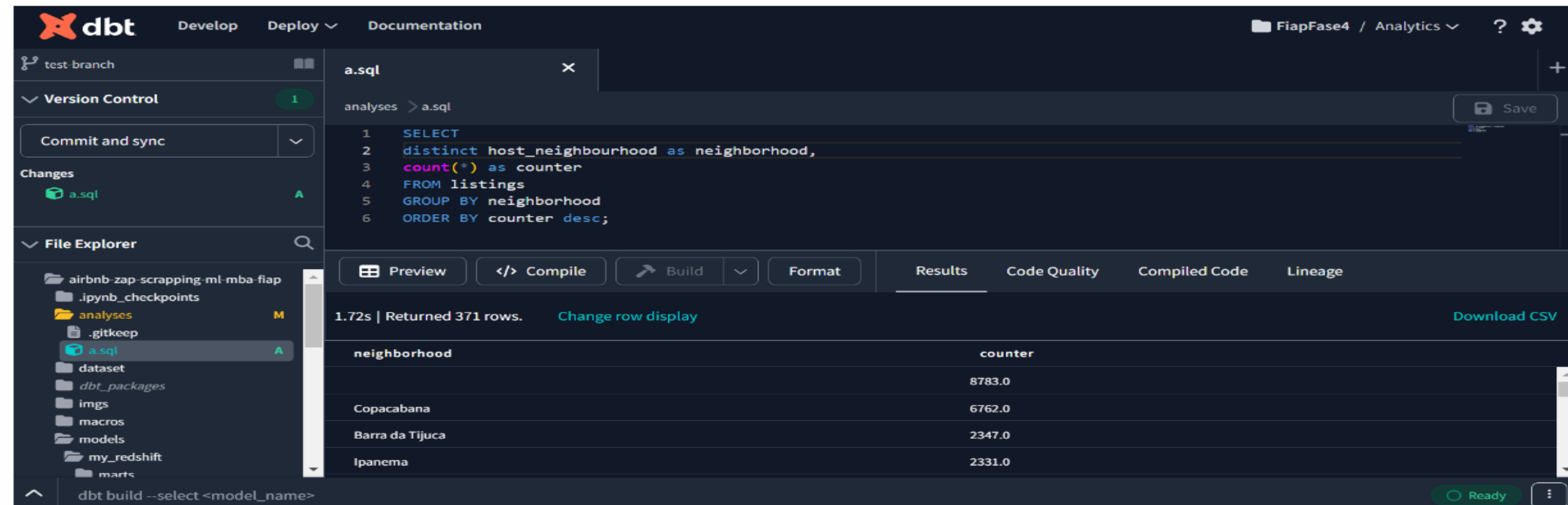
In the 'Lineage' view, a node '2+zap_data+2' is visible, and an 'Update Graph' button is present. At the bottom center, a purple box contains the text 'zap_data' next to a cube icon.

Análise de Dados Airbnb + Visualização no Metabase

Quantas acomodações existem num bairro e onde ficam?

a) How many accommodations are there in a neighborhood and where are they located?

There are 371 different neighbourhood values - including the null values, and Copacabana, Barra da Tijuca, and Ipanema lead the ranking.



The screenshot shows the dbt CLI interface. The top bar includes the dbt logo, navigation tabs (Develop, Deploy, Documentation), and the current project path (FiapFase4 / Analytics). The left sidebar shows the file explorer with the 'analyses' directory selected, containing a file named 'a.sql'. The main panel displays the SQL query in 'a.sql' and its execution results.

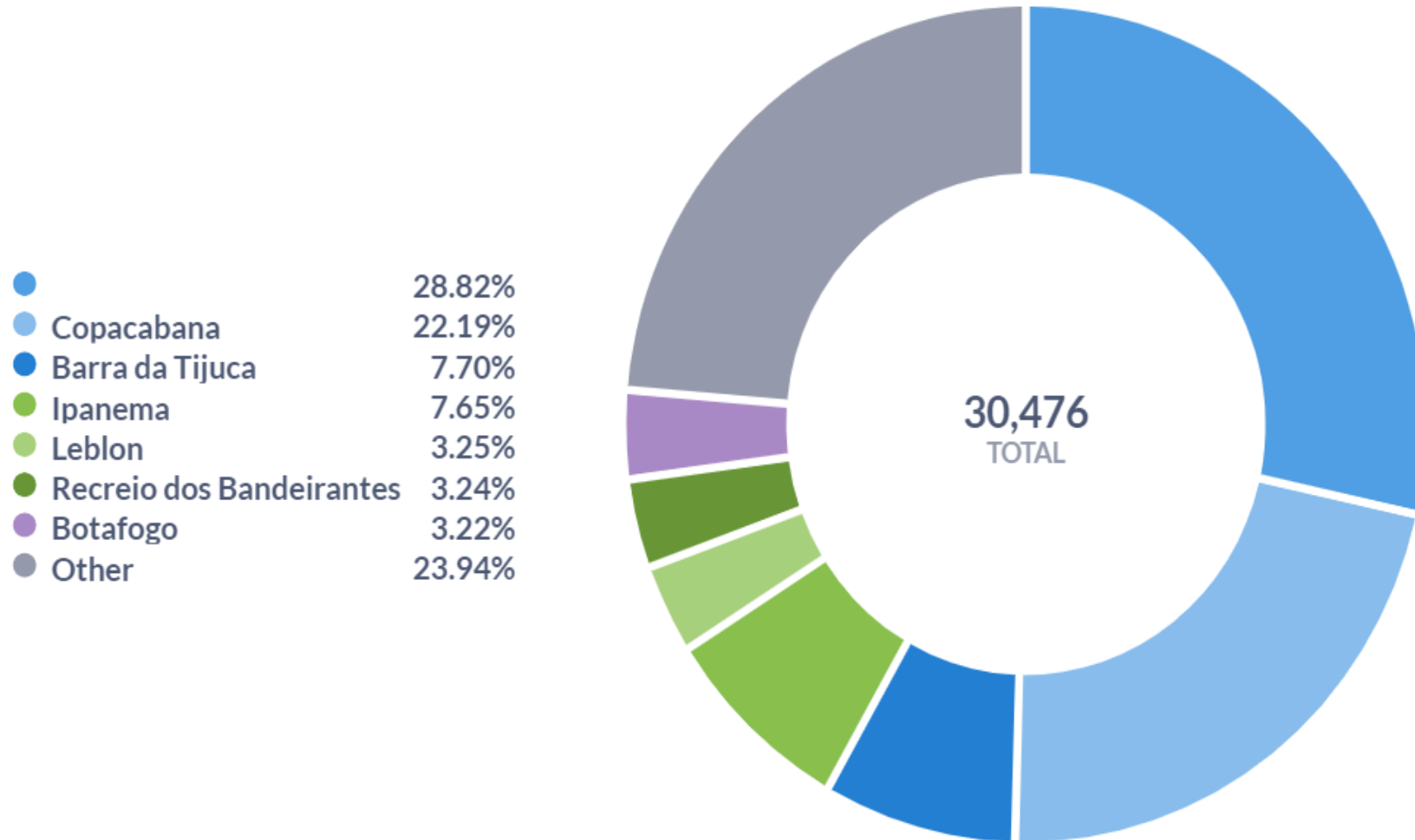
```
1 SELECT
2 distinct host_neighbourhood as neighborhood,
3 count(*) as counter
4 FROM listings
5 GROUP BY neighborhood
6 ORDER BY counter desc;
```

The results table shows the top neighborhoods by accommodation count:

neighborhood	counter
	8783.0
Copacabana	6762.0
Barra da Tijuca	2347.0
Ipanema	2331.0

The bottom status bar indicates the command 'dbt build --select <model_name>' and the status 'Ready'.

Quantas acomodações existem num bairro e onde ficam?



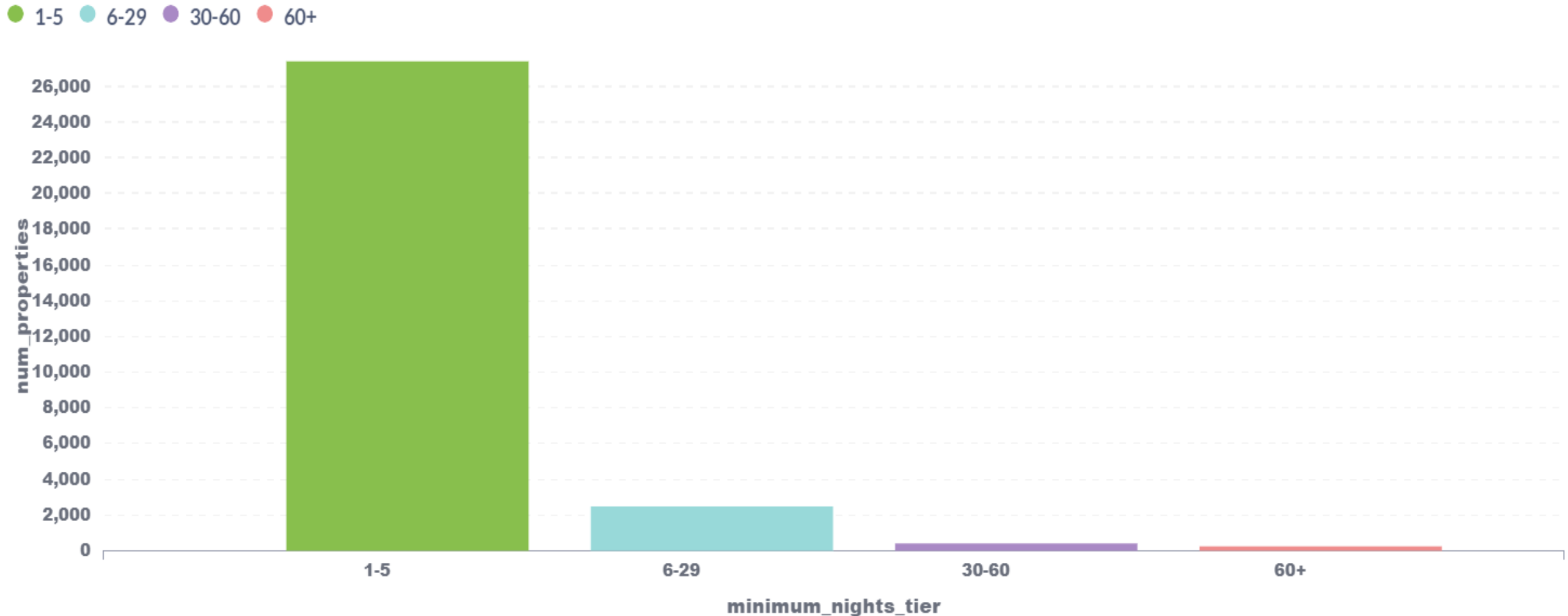
Quantas casas e apartamentos estão sendo alugados com frequência para turistas e não para residentes de longa duração?

b) How many houses and apartments are frequently being rented to tourists and not for long-term residents?

The great majority of the listings are for super short term rentals, with over 90% of the properties being available for rental with the number of minimum nights between 1-5 and almost 99% available for less than 30 nights.

minimum_nights_tier	num_properties
1-5	27390.0
6-29	2494.0
30-60	389.0
60+	203.0

Quantas casas e apartamentos estão sendo alugados com frequência para turistas e não para residentes de longa duração?



Quanto os hosts ganham alugando para turistas?

c) How much do hosts earn from renting to tourists?

Hosts earnings can vary based on multiple factors. However, considering the following assumptions:

- median price of \$350/night/person
- average number of reviews/month of 1.01
- average "acommodates" of 4, but likely to fill only half of that
- trend previous identified of short stays pointing to weekend stays (2 days)
- review ratio of 25% (1 out of 4 people that rent an airbnb leave a review)

We could calculate the average monthly income of a host by:

- Calculating the average rental: $350\$ \times 2 \text{ people} \times 2 \text{ days} = \$ 1.400$
- Calculating the number of rentals/month: $1/0,25 \times \$ 3.924 = \sim 6.400$

This number seems feasible, but can drastically vary based on the assumptions above and specific airbnb variables.

Quais hosts estão administrando uma empresa com várias listagens e onde estão?

d) Which hosts are managing a business with multiple listings and where are they located?

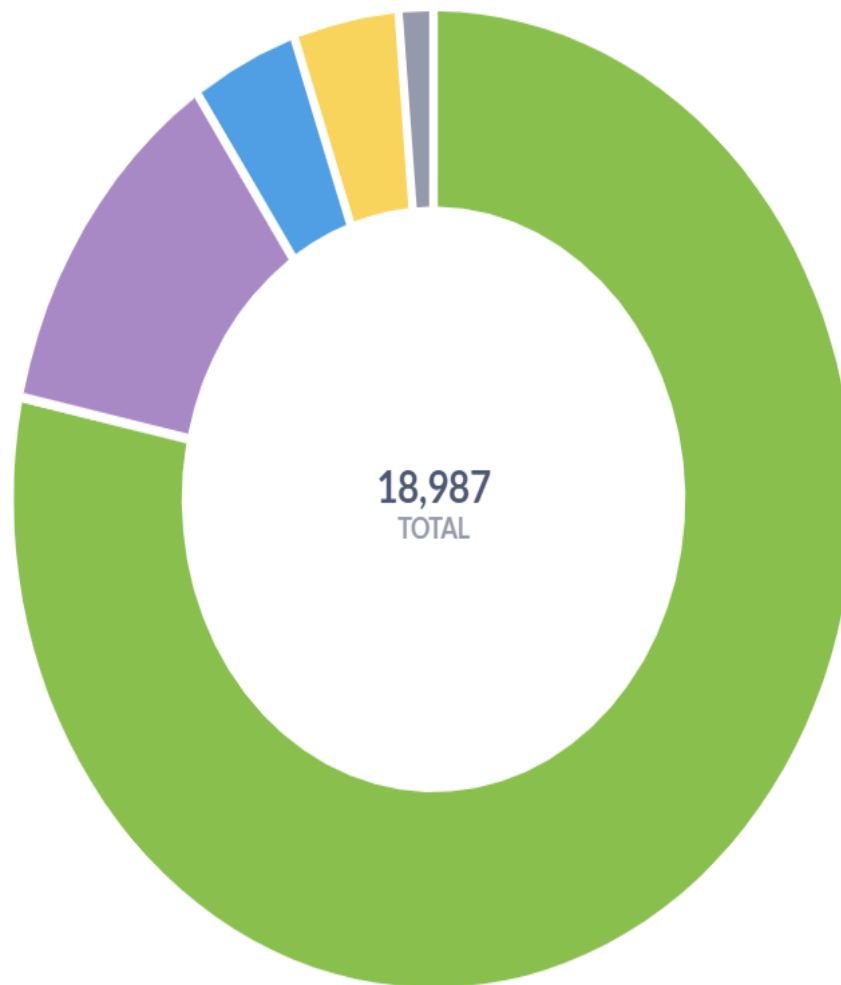
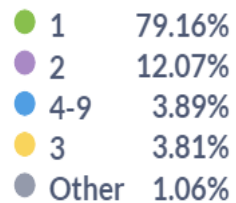
Almost 80% of the hosts have a single listing, and only 5% have more than 3 listings.

7.57s Returned 6 rows. Change row display	
num_listings_tier	num_hosts
1	15031.0
2	2292.0
4-9	738.0
3	724.0
10-49	186.0
50+	16.0

Investigating the hosts with highest number of listings, it is possible to identify that all of them concentrate properties in the same or few neighbourhoods inside noble city areas.

1.57s Returned 124 rows. Change row display		
host_id	neighbourhood_cleansed	num_listings
341887136.0	Ipanema	99.0
341887136.0	Leblon	62.0
341887136.0	Copacabana	28.0
331210726.0	Jacarepaguá	15.0
325956962.0	Centro	83.0
321818201.0	Copacabana	12.0

Quais hosts estão administrando uma empresa com várias listagens e onde estão?



Que tipo de acomodação é mais comum no Airbnb numa localidade?

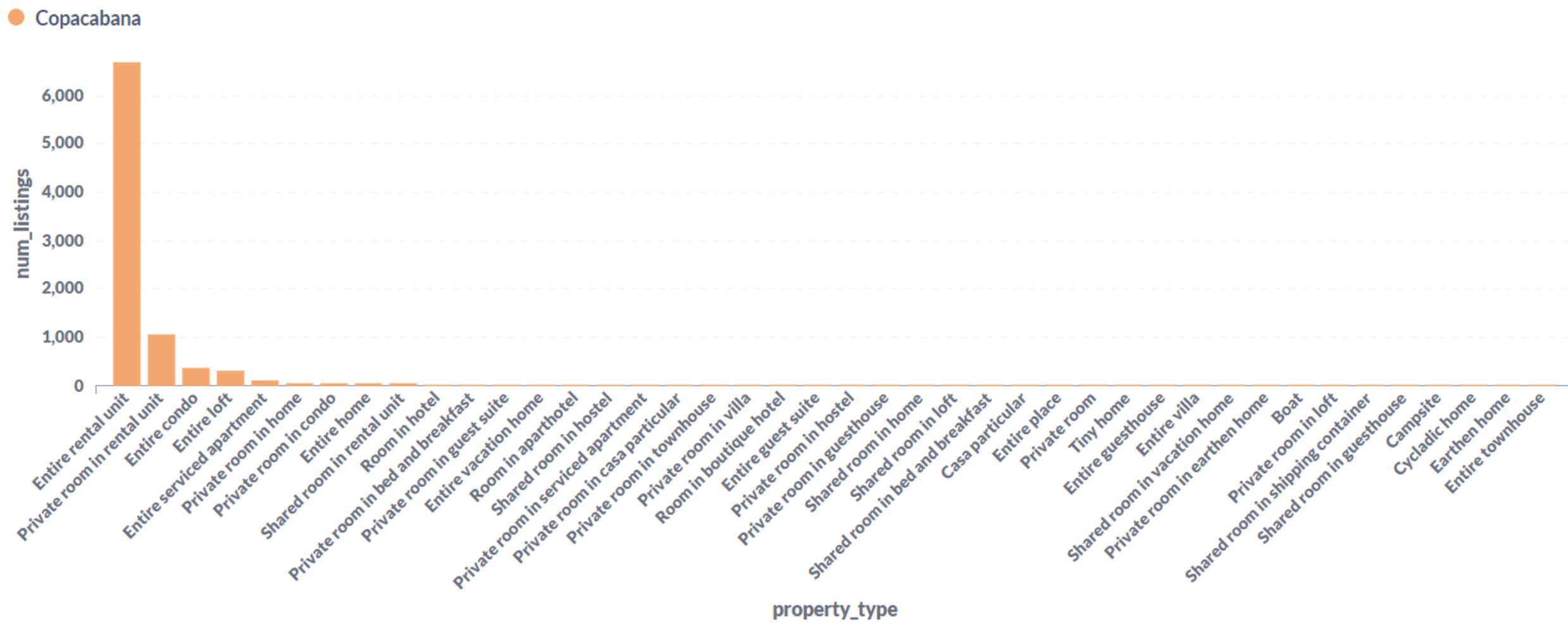
e) What type of accommodation is most common on Airbnb in a specific location?

For Copacabana, entire rental unit is by far the most common property type, followed by private room in rental unit.

6.80s | Returned 42 rows. [Change row display](#)

neighbourhood_cleansed	property_type	num_listings
Copacabana	Entire rental unit	6678.0
Copacabana	Private room in rental unit	1047.0
Copacabana	Entire condo	363.0
Copacabana	Entire loft	308.0
Copacabana	Entire serviced apartment	110.0
Copacabana	Private room in home	61.0

Que tipo de acomodação é mais comum no Airbnb numa localidade?



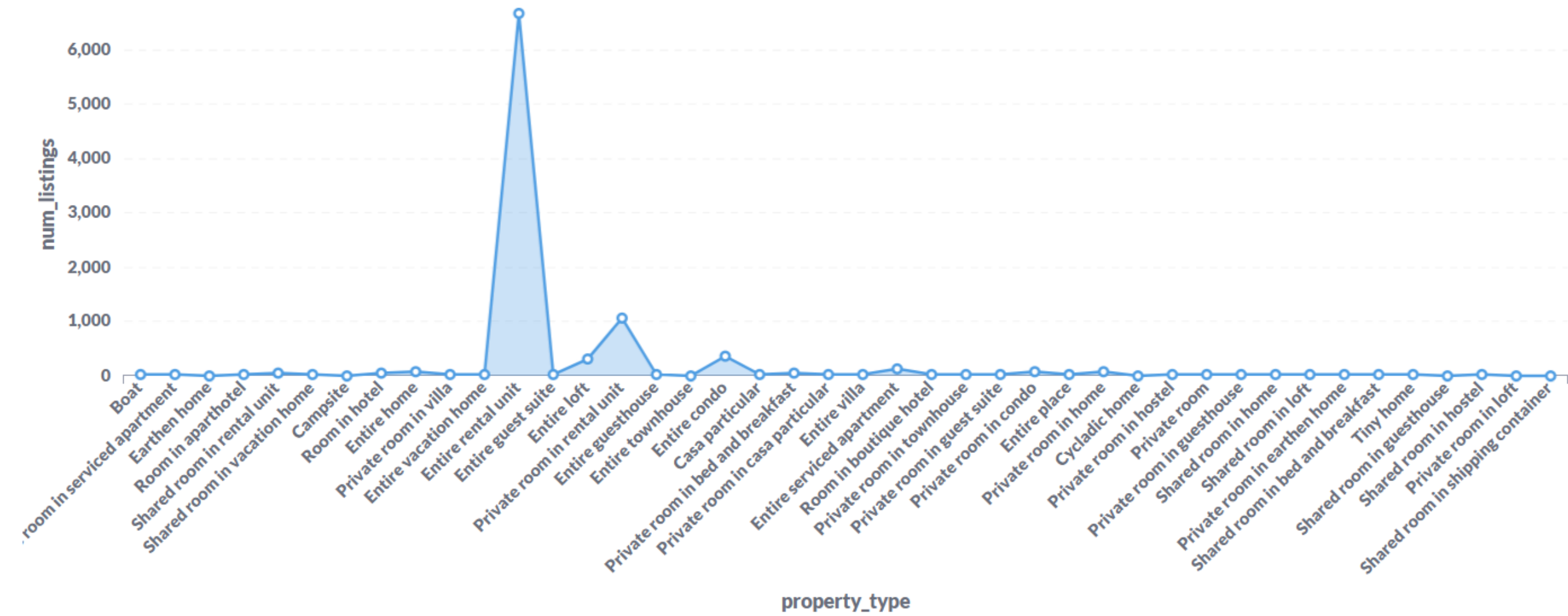
Qual é a diferença de preço entre os diferentes tipos de acomodações?

f) What is the price difference between different types of accommodations?

As it can be observed in the image below, for the neighbourhood of Copacabana there is a high variation between the price of different types of accommodations. Going from \$53 in a "Shared room in shipping container" and \$78 in a "Shared room in hostel" to \$6.250 for a boat. Entire rental unit, the most common property type has an average price of \$1.081.

neighbourhood_cleansed	property_type	avg_price	num_listings
Copacabana	Boat	6250.0	2.0
Copacabana	Private room in serviced apartment	4588.0	12.0
Copacabana	Earthen home	3203.0	1.0
Copacabana	Room in aparthotel	3072.0	18.0
Copacabana	Shared room in rental unit	2228.0	50.0
Copacabana	Shared room in vacation home	1800.0	2.0
Copacabana	Campsite	1600.0	1.0
Copacabana	Room in hotel	1342.0	33.0
Copacabana	Entire home	1198.0	54.0
Copacabana	Private room in villa	1185.0	8.0
Copacabana	Entire vacation home	1165.0	20.0

Qual é a diferença de preço entre os diferentes tipos de acomodações?



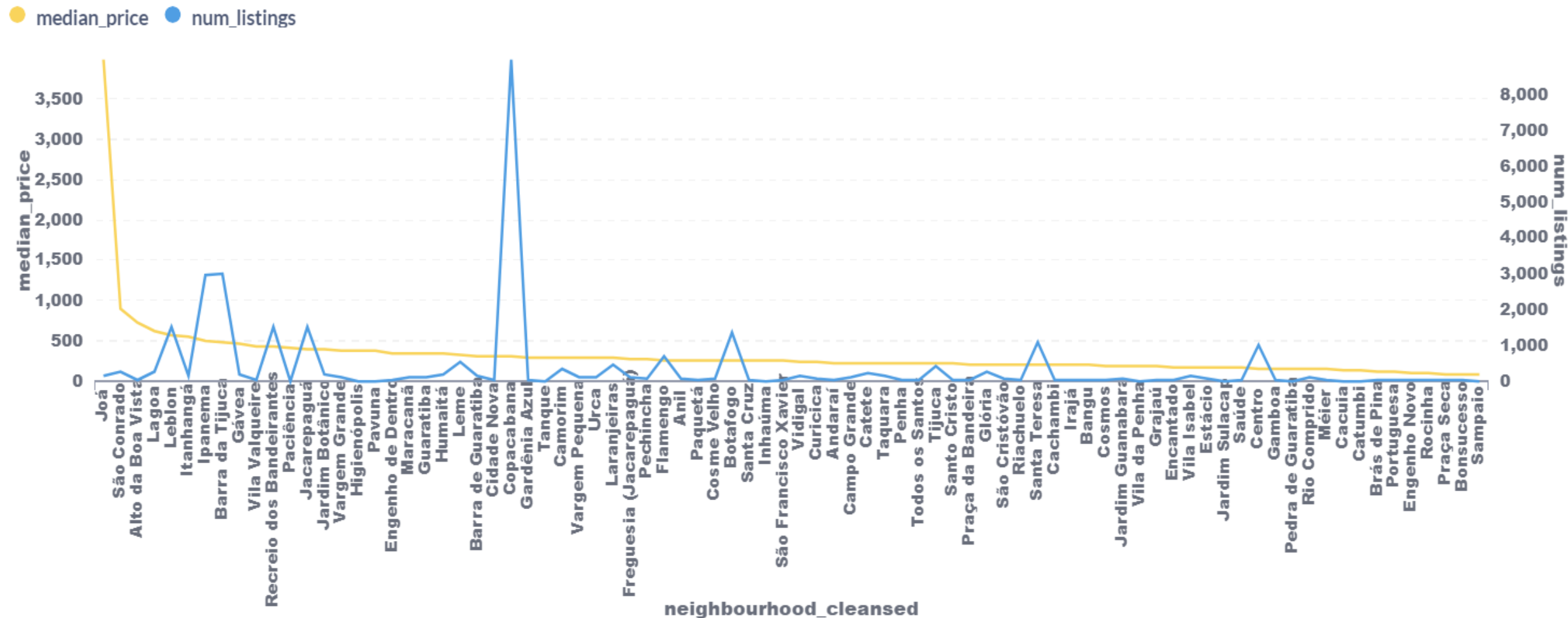
Analizando os dados do Airbnb...

g) What are the most expensive regions to stay in?

Using the median price instead of average to avoid outliers and filtering for neighbourhoods with more than 10 listings, Joá is by far the most expensive region followed by São Conrado, Alto da Boa Vista, Lagoa and Leblon.

neighbourhood_cleansed	median_price	num_listings
Joá	3992.0	135.0
São Conrado	900.0	258.0
Alto da Boa Vista	732.0	44.0
Lagoa	627.5	262.0
Leblon	575.0	1505.0
Itanhangá	558.5	152.0

Quais são as regiões mais caras para ficar?



Catálogo de Datos

Machine Learning

Machine Learning

Obrigado!

