SOLUTION SPRINT

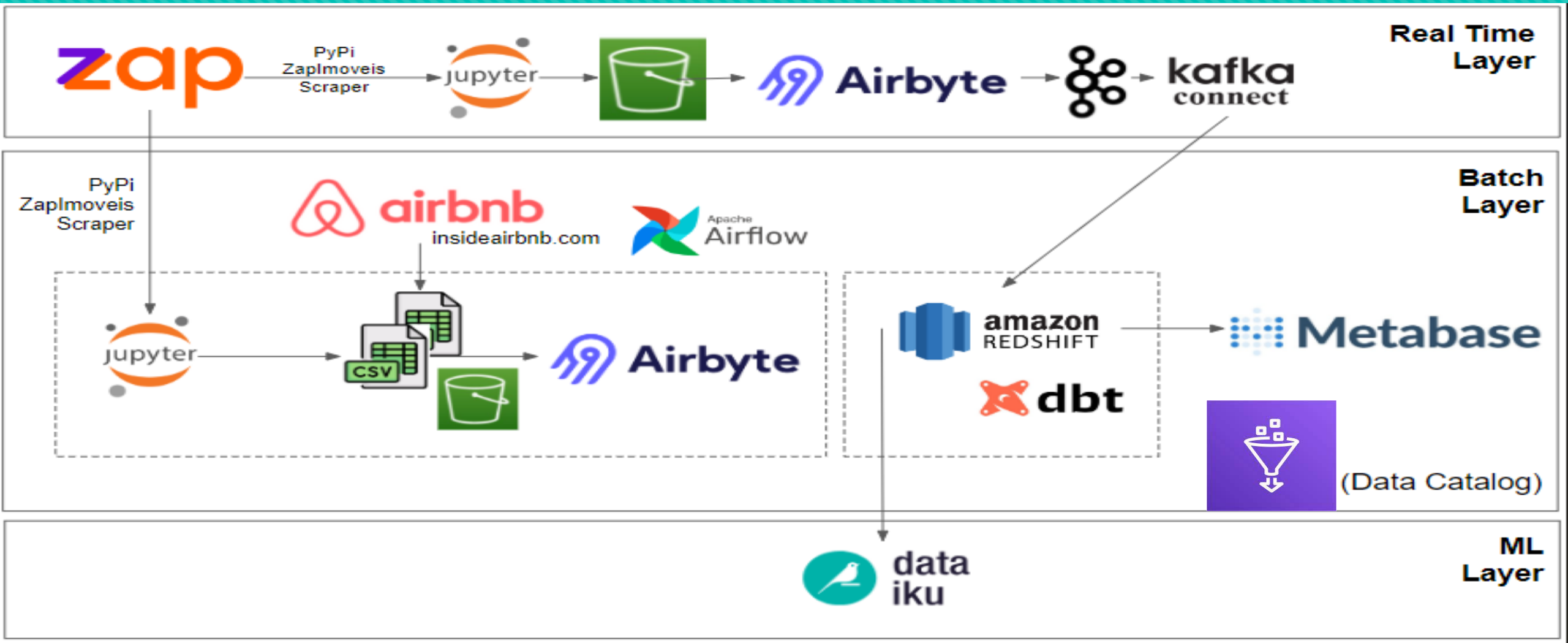
MBA ENGENHARIA DE DADOS – FASE 4

Equipe:
Marcelo Canabrava
Marina Coutinho

Descrição completa do projeto no github

https://github.com/mcanabrava/airbnb-zap-scrapping-ml-mba-fiap

Diagrama da Solução



Scrapping sugerido + proprietário

zap_bs4.csv	fix scrapper area
zap_imoveis_rj.csv	add scrapper code

In [8]:

```
zap_bs4.head(100)
```

Out[8]:

	neighborhood	city	description	price	condominio	iptu	area	bedrooms	suite	bathrooms	gara
0	Centro	Rio de Janeiro	UM CHARMOSO EDIFÍCIO HISTÓRICO, COMPLETAMENTE ...	367000	0	0	28	1	0	1	
1	Centro	Rio de Janeiro	O CENTRO DO RIO ESTÁ SEREQUALIFICANDO. MORAR N...	290000	0	0	32	1	0	1	
2	Centro	Rio de Janeiro	Economia criativa e inovação. Esses são os pil...	490000	0	0	33	1	0	1	
3	Copacabana	Rio de Janeiro	CONJUGADÃO REFORMADO EM RUA NOBRE DE COPACABAN...	450000	590	900	30	1	0	1	
4	Ipanema	Rio de Janeiro	Tenho o prazer de apresentar a você este belís...	2850000	2700	650	80	2	1	3	
...	...	...	...	...	...	...	...	...	...	...	

Armazenamento com upload via boto3 para o S3

zap

zap_bs4.csv	fix scrapper area
zap_imoveis_rj.csv	add scrapper code

airbnb
insideairbnb.com

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

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

FIAP PROJECT

Connections

Sources

Destinations

Builder

Billing

Resources

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

Resources Info

Select database Info
To view schemas, select a database.
dwhdatabase

Select schema Info
To view tables, select a schema.
public

Filter tables

1

- _airbyte_raw_s3_to_rs_zap_p...
- _airbyte_raw_s3_to_rs_zap
- s3_to_rs_zap
  - area
  - city
  - link
  - garage
  - description
  - _ab_source_file_last_modified
  - bathrooms
  - _ab_source_file_url

Status Connected | database dwhdatabase | user dwhuser | Change connection

Query 1

```
SELECT
    *
FROM s3_to_rs_zap
LIMIT 10;
```

Run | Save | Schedule | Clear | Send feedback

Query results | Table details

Query 654

Execution | Data | Visualize

Completed, started on August 09, 2023 at 20:38:10
ELAPSED TIME: 00 m 02 s

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 97m² 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

dbt | Develop | Deploy | Documentation | FiapFase4 / Analytics

test-branch | Change branch

Version Control

Create a pull request on ...

File Explorer

- my_redshift
  - marts
  - staging
    - calendar.sql
    - listing_samples.sql
    - listings.sql
    - review_samples.sql
    - schema.yml
    - zap_data.sql
  - seeds
  - snapshots
  - target

calendar.sql | zap_data.sql

models > my_redshift > staging > zap_data.sql

Save

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

Preview | Compile | Build | Format | Results | Code Quality | Compiled Code | Lineage

2+zap_data+2 | Update Graph

zap_data

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.

dbt | Develop | Deploy | Documentation | FiapFase4 / Analytics

test-branch

Version Control 1

Commit and sync

Changes

a.sql A

File Explorer

airbnb-zap-scrapping-ml-mba-fiap
- .ipynb_checkpoints
- analyses M
  - .gitkeep
  - a.sql A
- dataset
- dbt_packages
- imgs
- macros
- models
  - my_redshift
    - marts

a.sql

analyses > a.sql

Save

```sql
SELECT
distinct host_neighbourhood as neighborhood,
count(*) as counter
FROM listings
GROUP BY neighborhood
ORDER BY counter desc;
```

Preview | Compile | Build | Format | Results | Code Quality | Compiled Code | Lineage

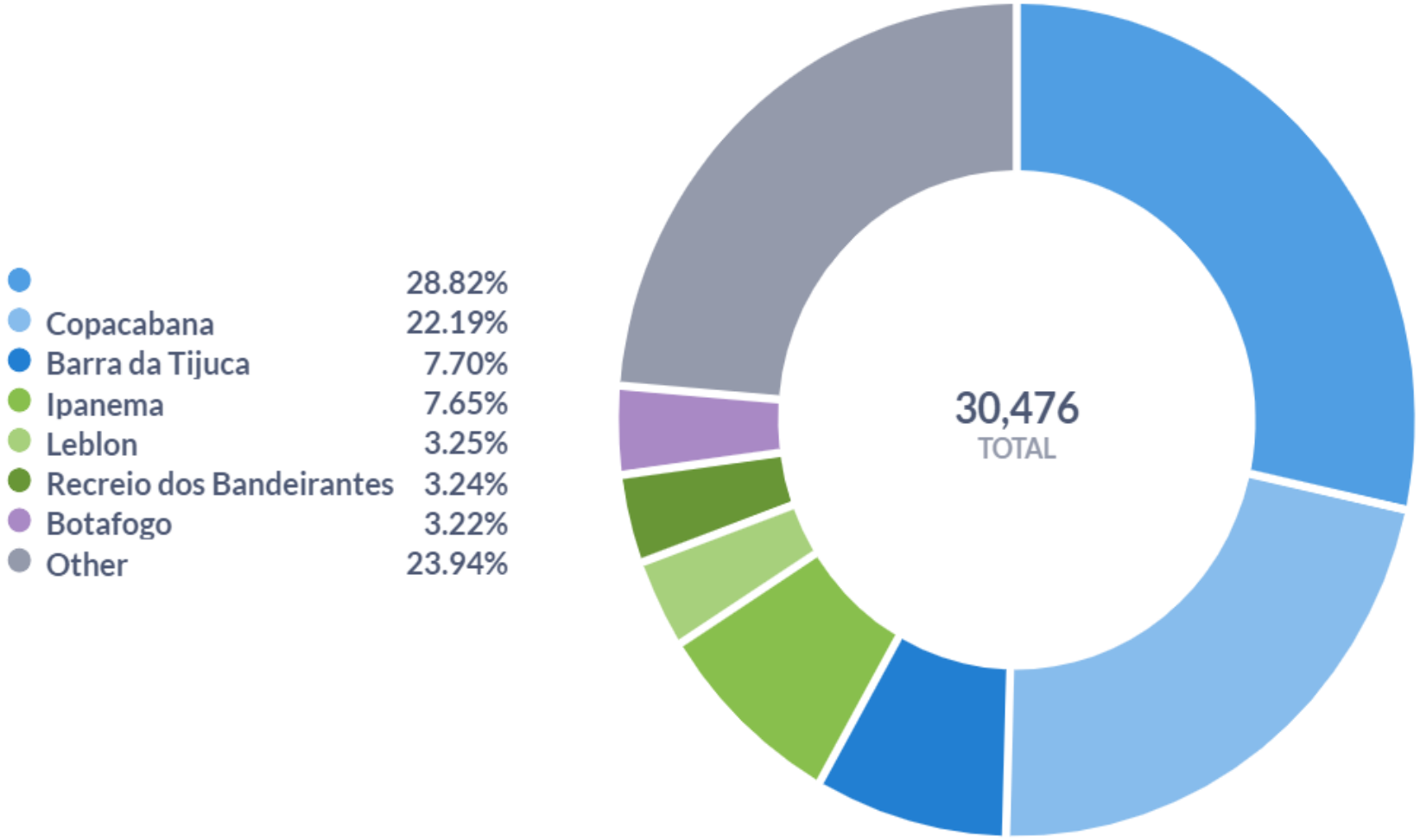
1.72s | Returned 371 rows. Change row display Download CSV

neighborhood	counter
	8783.0
Copacabana	6762.0
Barra da Tijuca	2347.0
Ipanema	2331.0

dbt build --select <model_name>

Ready

Quantas acomodações existem num bairro e onde ficam?



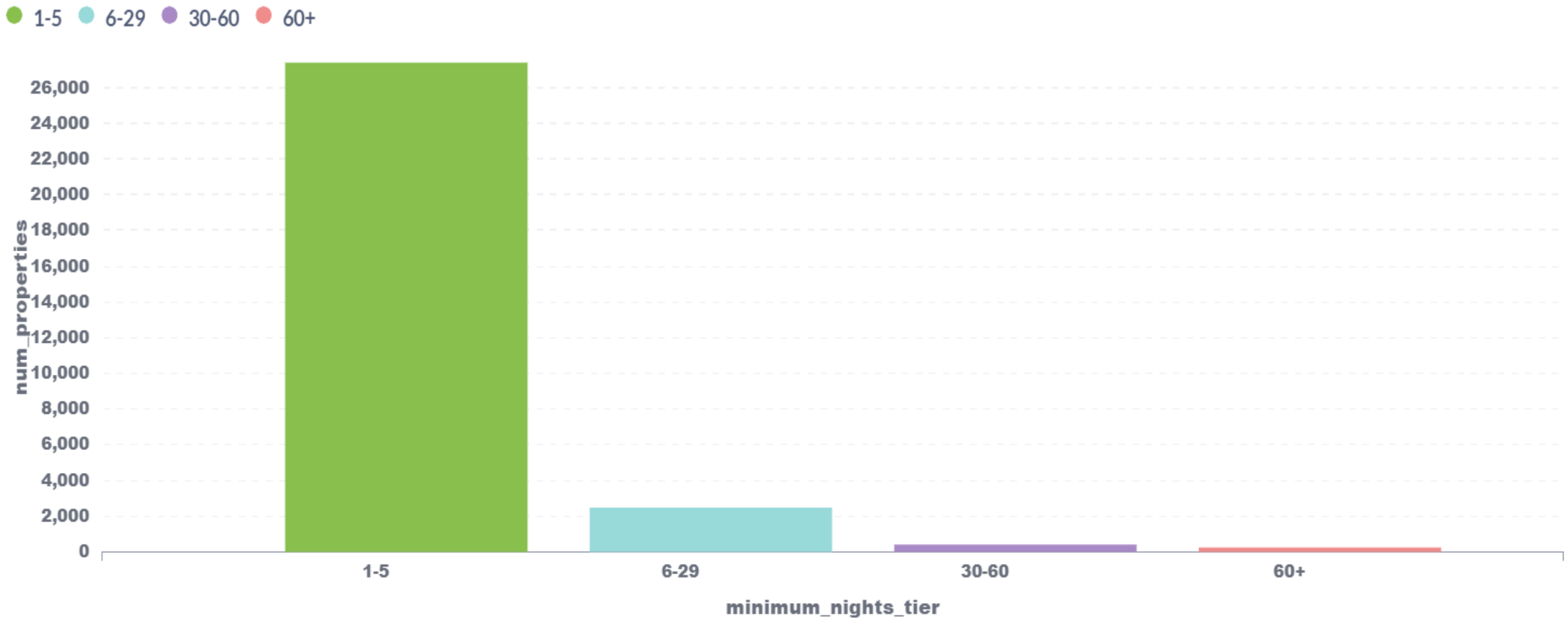
Quantas casas e apartamentos estão sendo alugados com frequência para turistas e não para residentes de longa duração?

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$ x 2 people x 2 days = $ 1.400
- Calculating the number of rentals/month: 1/0,25 x $ 3.924 = ~ 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

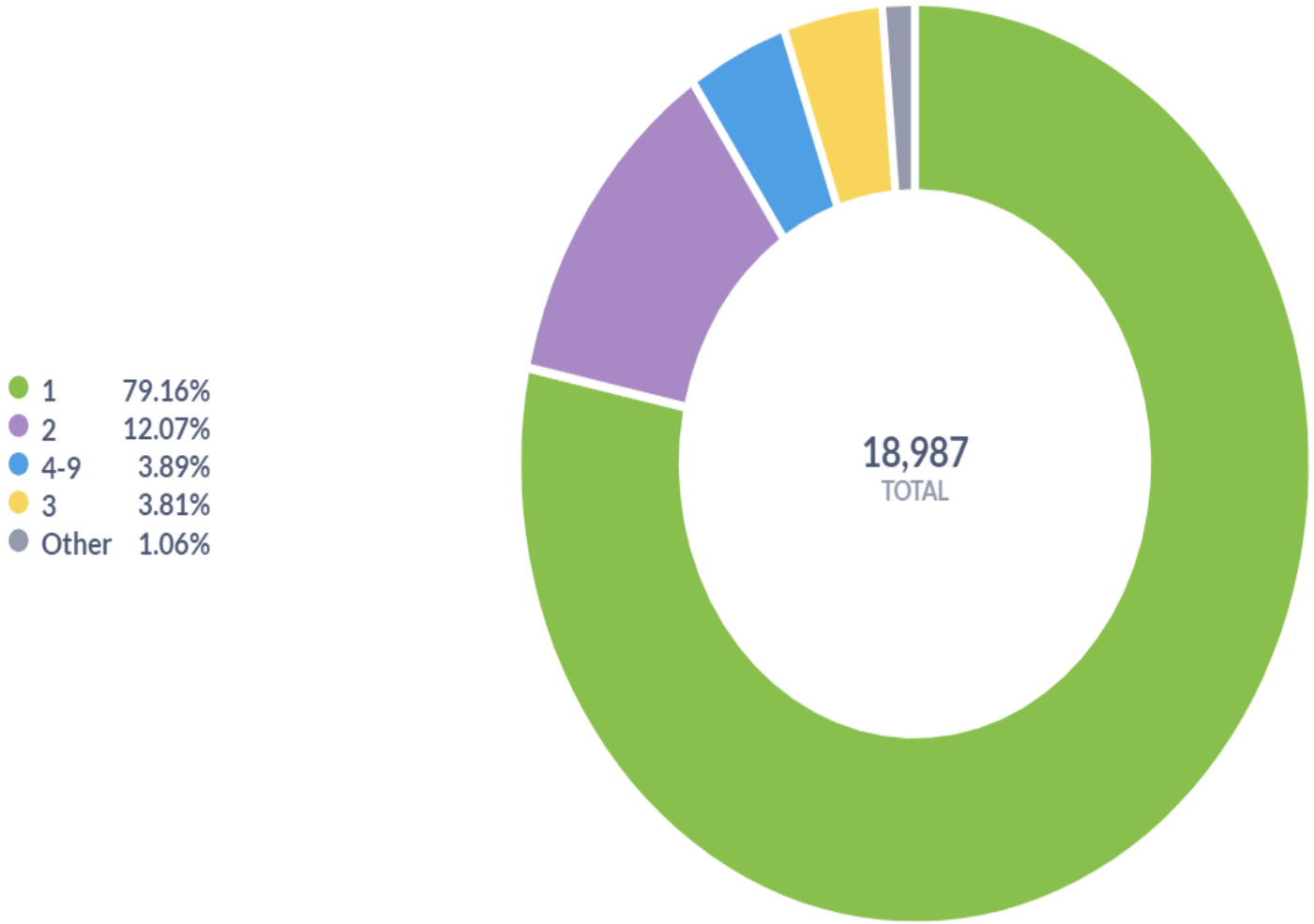
num_listings_tier	num_hosts
1	15031.0
2	2292.0
4-9	738.0
3	724.0
10-49	186.0
50+	16.0

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?

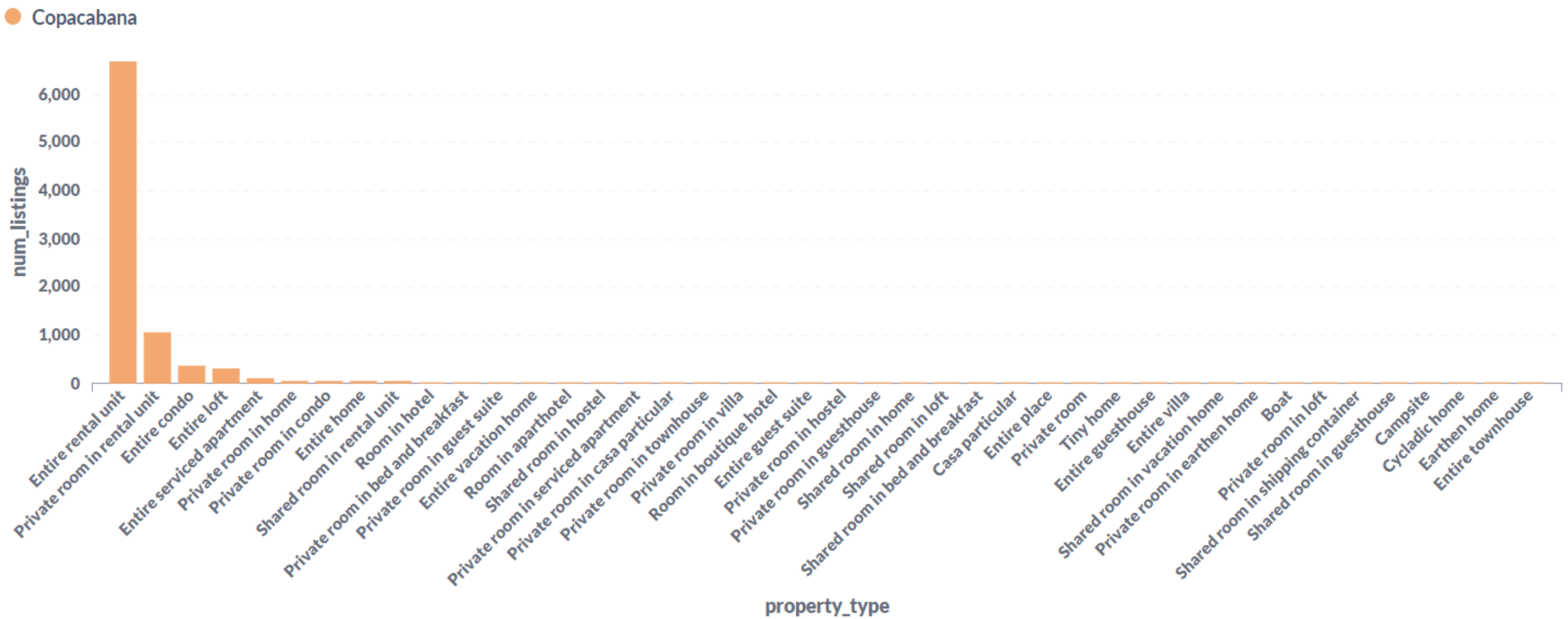
e) What type of accommodation is most common on Airbnb in a specific location?

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?



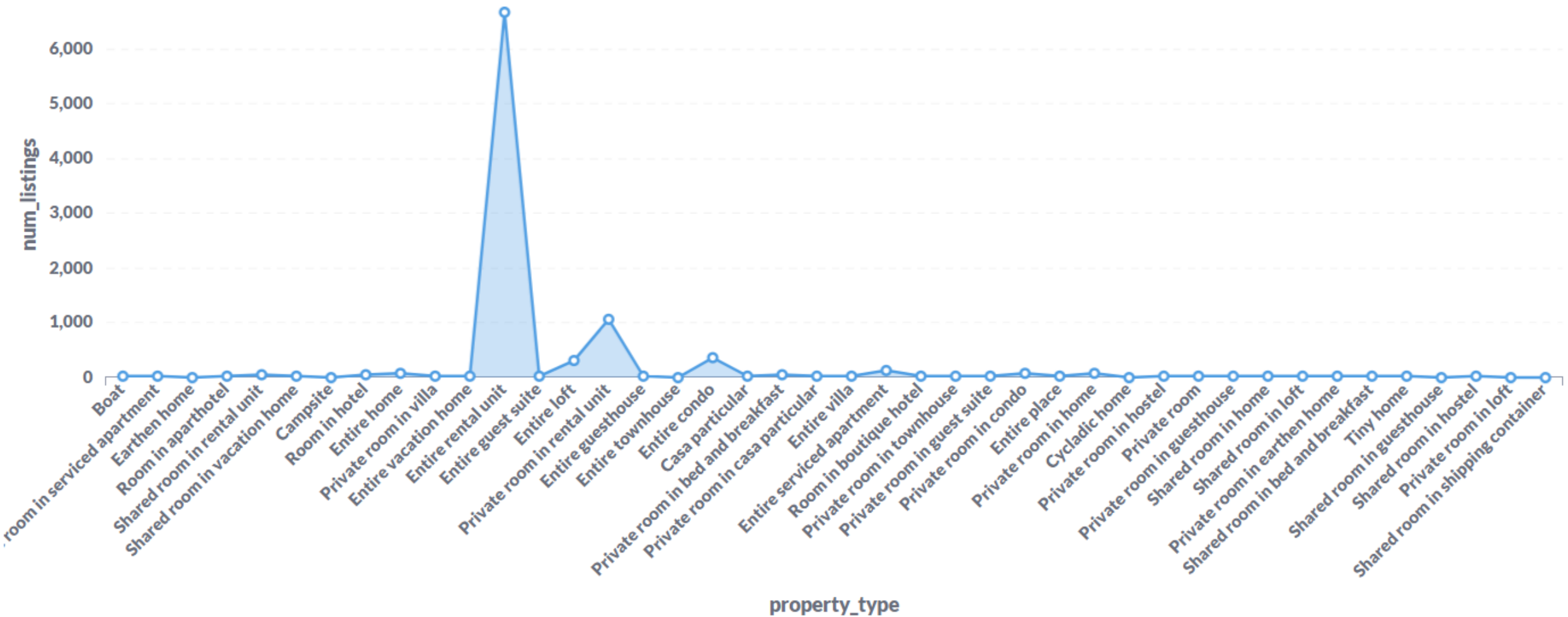
Qual é a diferença de preço entre os diferentes tipos de acomodações?

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?



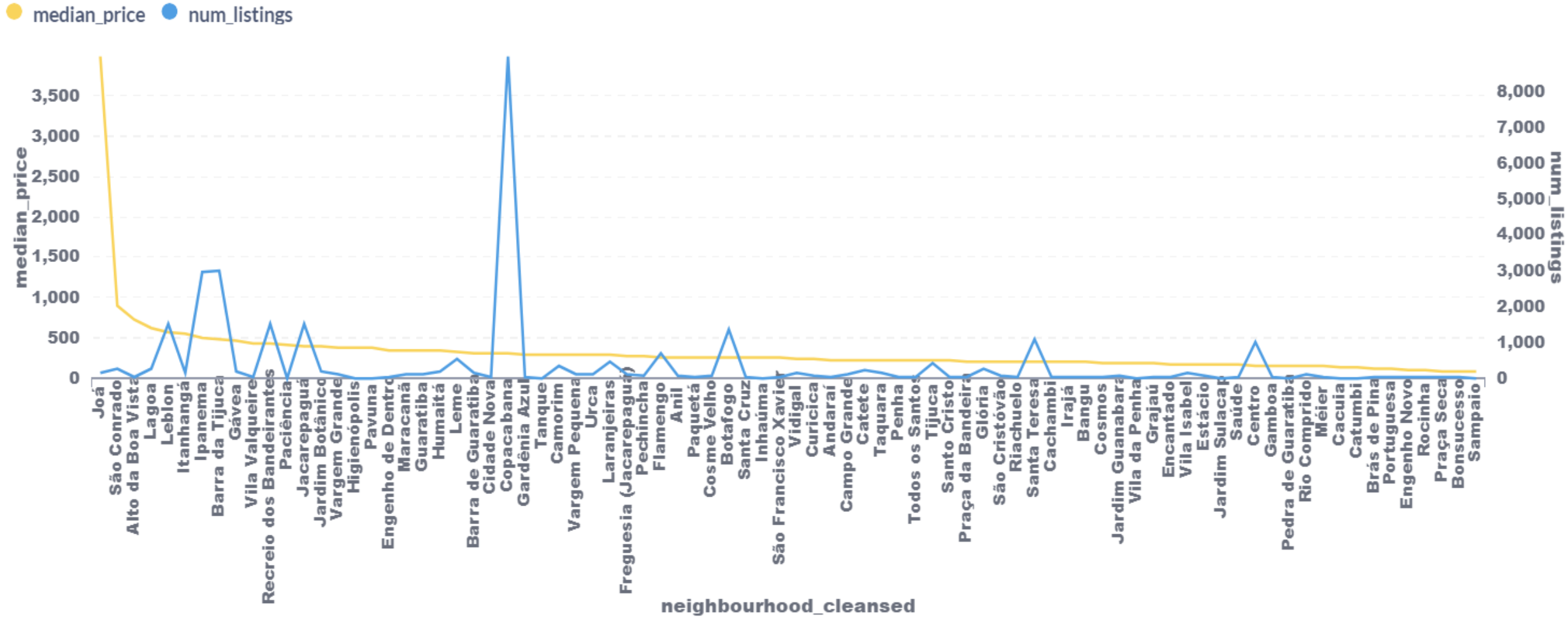
Analisando os dados do Airbnb...

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 Dados com AWS Glue

redshift-crawler

Last updated (UTC)
September 11, 2023 at 21:18:15

Run crawler | Edit | Delete

Crawler properties

Name	IAM role	Database	State
redshift-crawler	glue-crawler-role	fiap-fase4	READY
Description	**Security configuration**	**Table prefix**	
-	-	-	

Advanced settings

Inherit schema from table	Schema updates in the data store	Object deletion in the data store	Create Partition Index
False	Update the table definition in the data catalog	Mark the table as deprecated in the data catalog.	False

Crawler runs | Schedule | Data sources | Classifiers | Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Stop run | View CloudWatch logs | View run details

Filter data | Filter by a date and time range

1

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
September 4, 2023 at 18:32:22	September 4, 2023 at 18:35:14	02 min 51 s	Completed	0.272	14 table changes, 0 partition changes

Crawler do AWS Glue para pegar as tabelas do Redshift e criar o catálogo de dados

Catálogo de Dados com AWS Glue

Tables (7)
View and manage all available tables.

Last updated (UTC)
September 11, 2023 at 21:18:45

Delete | Add tables using crawler | Add table

Filter tables

1

Name	Database	Location	Classification	Deprecated	View data	Data quality
dwhdatabase_public__	fiap-fase4	dwhdatabase.public._a	redshift	-	-	View data quality
dwhdatabase_public__	fiap-fase4	dwhdatabase.public._a	redshift	-	-	View data quality
dwhdatabase_public_c	fiap-fase4	dwhdatabase.public.cal	redshift	-	-	View data quality
dwhdatabase_public_li	fiap-fase4	dwhdatabase.public.list	redshift	-	-	View data quality
dwhdatabase_public_li	fiap-fase4	dwhdatabase.public.list	redshift	-	-	View data quality
dwhdatabase_public_re	fiap-fase4	dwhdatabase.public.rev	redshift	-	-	View data quality
dwhdatabase_public_s	fiap-fase4	dwhdatabase.public.s3	redshift	-	-	View data quality

Tabelas de dados criadas a partir do Redshift pelo Crawler.

Schema (7)
View and manage the table schema.

Edit schema as JSON | Edit schema

Filter schemas

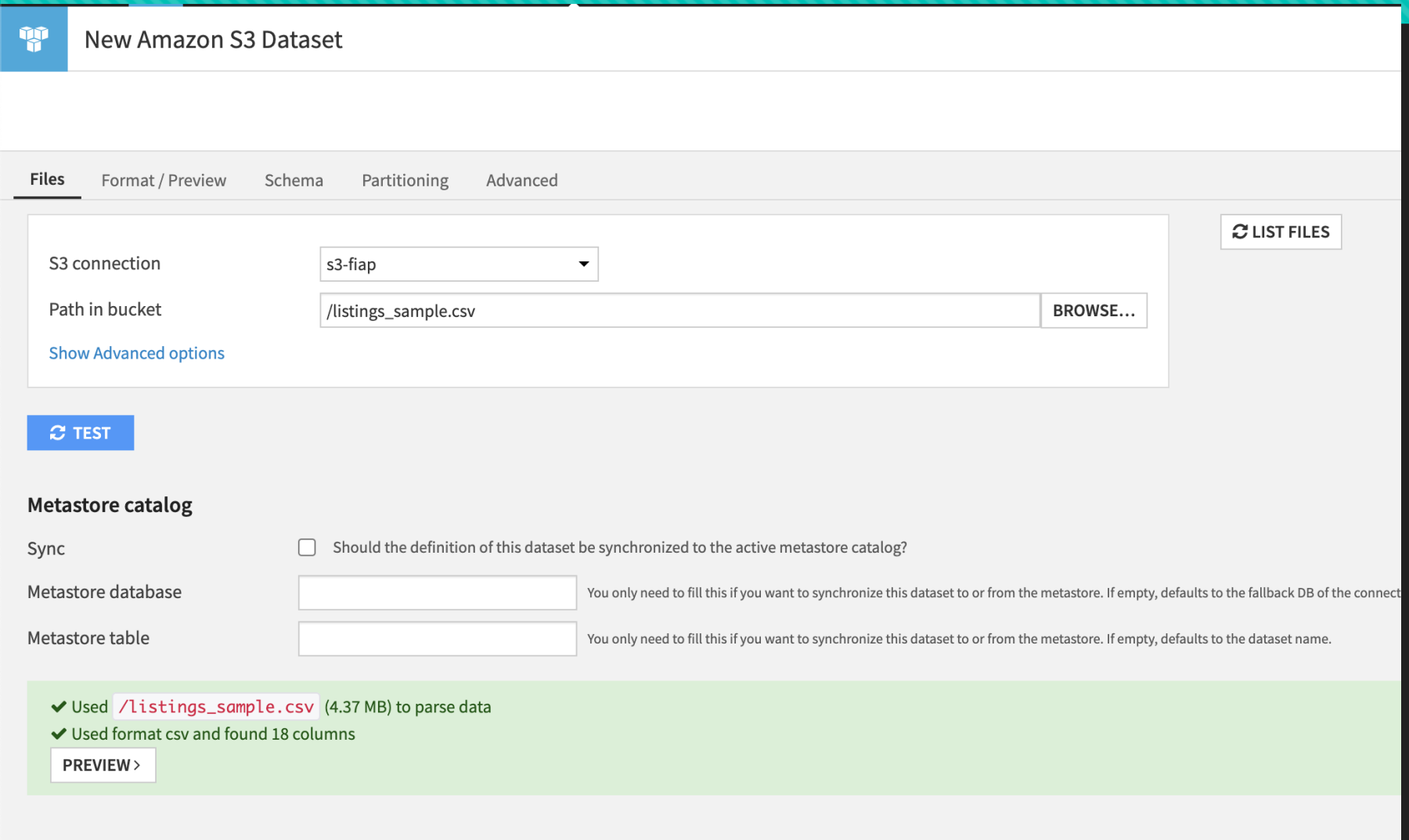
1

#	Column name	Data type	Partition key	Comment
1	date	date	-	-
2	listing_id	bigint	-	-
3	minimum_nights	string	-	-
4	price	decimal(18,0)	-	-
5	adjusted_price	decimal(18,0)	-	-
6	available	boolean	-	-
7	maximum_nights	string	-	-

Schema com os dados catalogados (exemplo de Uma tabela)

Machine Learning

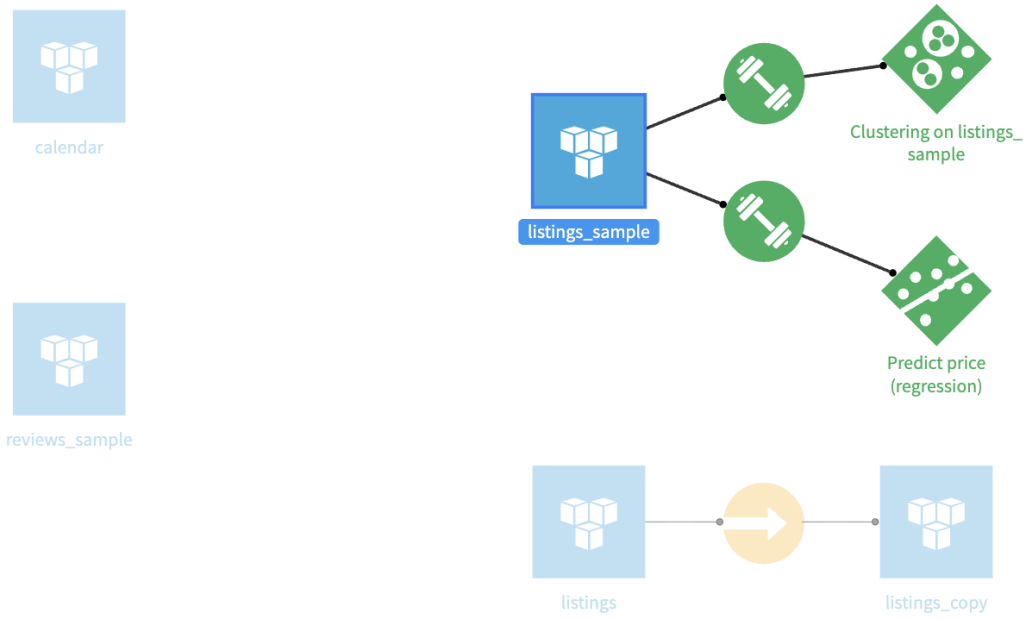
Machine Learning - Datalku



Datasets criados a partir dos CSVs armazenados no Amazon S3

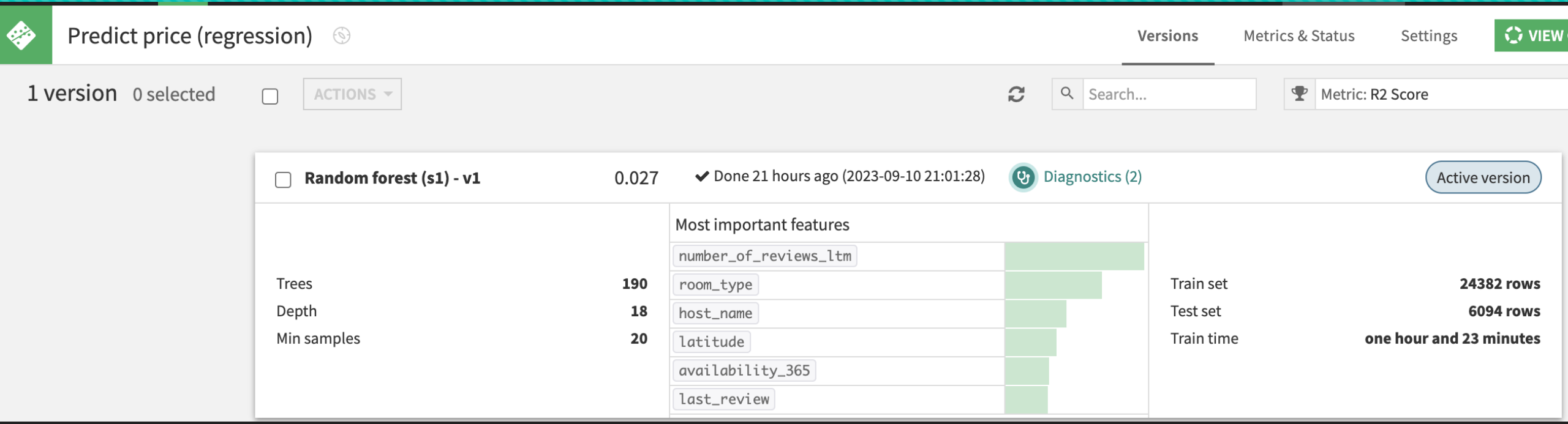
Machine Learning - Dataiku

3 recipes **5** datasets **2** models



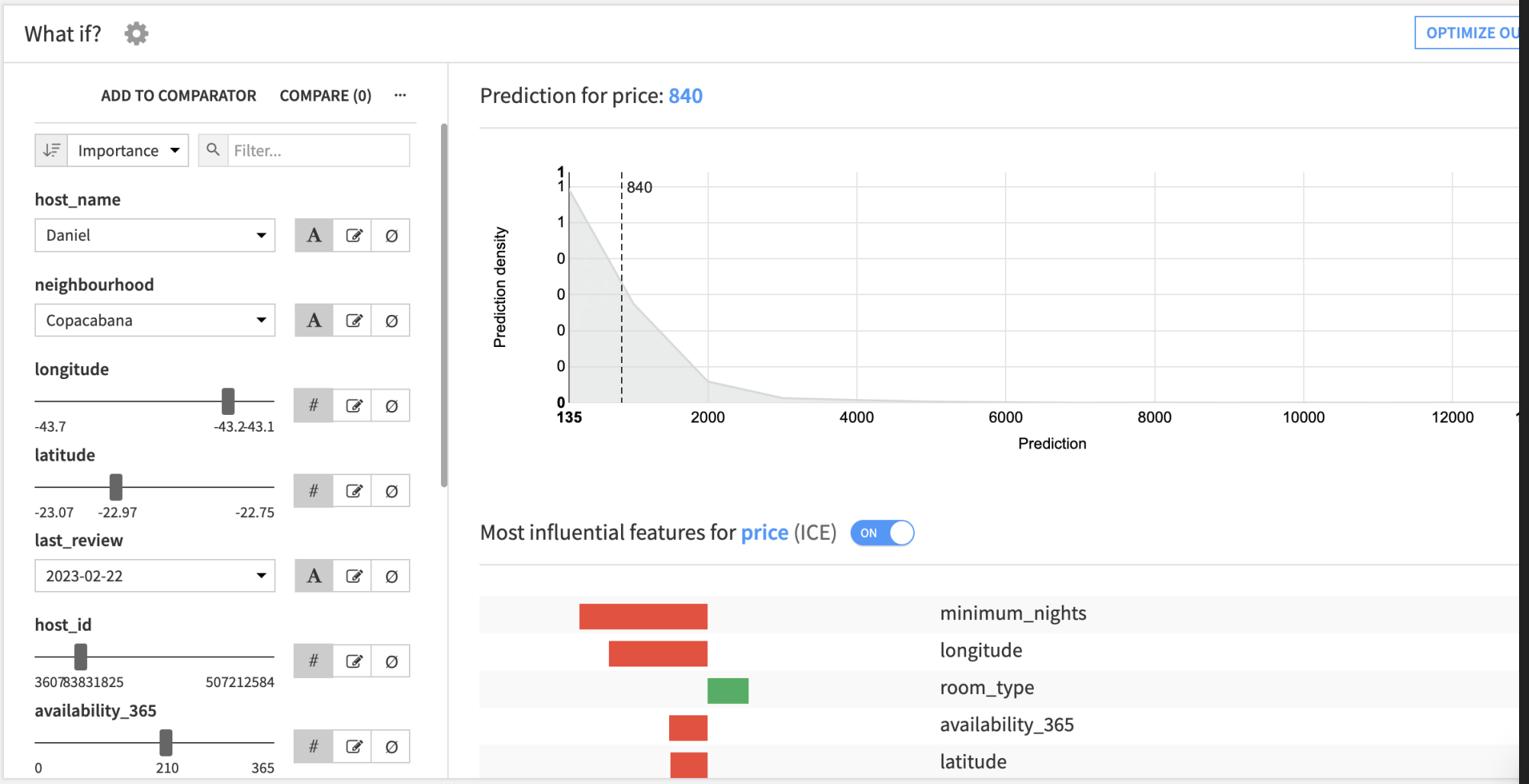
Listing_samples foi o dataset escolhido para rodar o ML para predição de texto, usando algoritmo de regressão do próprio Dataiku

Machine Learning - Datalku



Aqui, o algoritmo rankou os atributos mais influentes no preço

Machine Learning - Datalku



Baseado no treinamento do algoritmo, conseguimos escolher as variáveis e ver como as combinaçòes podem influenciar no preço da hospedagem, aqui, como seria por exemplo um Airbnb em Copacabana recém avaliado pode ajudar na valorização do preço de locação

Obrigado!

Marcelo Canabrava
Marina Medeiros