Seazone Challenge

Data Scientist

Candidate: André Padilha

Date: December 30th, 2022

1. Introduction

A total of 5 datasets were provided, containing Airbnb and Vivareal listings. Then 5 questions were made:

1. What is the best property profile to invest in the city?
2. Which is the best location in the city in terms of revenue?
3. What are the characteristics and reasons for the best revenues in the city?
4. We would like to build a building of 50 apartments in the city, where should we build it and how should the apartments be designed in order to be a great investment?
5. How much will be the return on investment of this building in the years 2024, 2025 and 2026?

The challenge was solved using Python and several libraries of Data Science. A preliminary analysis to explore and analyze data was done using Jupyter Notebooks, but the final solution is completely apart from it.

I have divided this project in 5 parts:

- Extract, Transform, Load (ETL):
  - One of the main challenges consisted of cleaning poorly formatted data and creating one-hot-encodings. Several steps were taken to address this issue and will be described further below.
  - One of the files had 4.6 GB, so it was necessary to use something beyond Pandas to load it. I chose Dask.
  - Some data were duplicated across files, and the Case description was used to guide which one to choose from.
  - Geohashes were used to divide cluster listings in the same area.
  - Some assumptions were made and detailed described further below.
- Real estate development cost estimate:
  - A custom Excel Sheet was used to input data and estimate costs to develop and construct the building described on Question 4.
- Data Analysis:
  - Several methods were created to plot data, allowing easy modifications of what and how to plot it.
  - All questions are answered here
- Future work and feedback:
  - Ideas to improve the analysis and challenge feedback.
- Questions answers
  - All questions are quickly answered here again.

1.1. Terms definitions:

As it is said in the case description, terms like “profile”, “characteristics” and “location” are not defined. I would like to dedicate this section to define them, as they will guide the rest of the analysis.

The term “best” will be defined on each question, since it varies depending on context.

1.1.1. Property profile:

A property **profile** is defined as a set of attributes that define a certain property, making it possible to group it with other properties sharing the same profile. Not necessarily all characteristics need to be used at once to group properties, that is, it is completely possible to use only one attribute to group similar properties (as in “Type of property”).

Within this case solution, property profile attributes are:

- Location
- Type of property: for example apartment (whole), house (whole), apartment (one room), house (one room), Trailer, Camping, and so on.
- Number of bedrooms
- Number of beds
- Number of bathrooms
- Parking space
- Pool
- Gym
- Amenities
- Ocean view

1.1.2. Characteristics:

A given Airbnb listing has a set of characteristics (also called **features**) that are relevant (or not) to guests. It includes the property profile.

Some of them can be added if the Property Owner decides to, but others cannot. Other characteristics besides the Property profile are:

- Wifi, SmartTVs and other entertainment systems
- Number of guests allowed
- Appliances
- Furnitures
- Pool
- Air conditioning
- Pet-friendly
- Children-friendly

- Check-in type: host, self check-in etc.
- Kitchen essentials
- Beach essentials
- Bedding

1.1.3. Location:

Using the locations[1] provided in the datasets I have grouped listings using **geohash**.

"*Geohash is a public domain geocode system invented in 2008 by Gustavo Niemeyer which encodes a geographic location into a short string of letters and digits*" (Wikipedia).

It divides Earth in consecutive rectangular cells, with the geohash length being directly proportional to its precision.

A given point (lat/lon) can be encoded in a geohash, meaning that it is inside the area described by the geohash.

[1] Listing locations are not always precise. Some of the reasons include Real Estate agents being afraid of losing clients if they provide the complete address.

For example, geohash “6” encodes most of South America.

2. Extract, Transform, Load (ETL)

2.1. Airbnb Listings

I have begun with “Details_Data.csv”, importing it with Pandas. I have called it **airbnb_listings**.

airbnb_listings is a Pandas Dataframe containing entries from “Details_Data.csv” and it contains entries describing Airbnb listings **characteristics** scraped from the website.

2.1.1. One-hot-encoding

Besides usual data cleaning processes (transforming strings to Datetime format, substituting commas with dots), I had to create one-hot-encodings of “safety_features”, “house_rules” and “amenities”.

They were all formatted as string, and it was supposed to be simple process, but some entries had formatting problems for example:

```
,\,E,x,t,i,n,t,o,r, ,d,e, ,i,n,c,\\,u,0,0,e,a,n,d,i,o,\,,, ,\,D,e,t,e,c,t,o,r, ,d,e, ,f,u,m,a,\\,u,0,0,e,7,a,\,
```

I created a method to remove unwanted characters and create a comma separable string that could be encoded. For the above example, the result is:

```
Extintor de incêndio,Detector de fumaça
```

Then it could be simply encoded using “Series.str.get_dummies”.

2.1.2. Dropping duplicates

There were many duplicate entries describing the same Airbnb listing. This was due to the fact that the scraper executed on different days.

To use only the most up to date data of each listing, it was sorted based on the day it was scraped, and older entries were dropped.

There’s a caveat though: some features, like “min_nights”, had no value (<*NA*>) in the latest scraped entry. I analyzed it to see if it was <*NA*> only on the latest scraped entry, but it was actually on several entries besides the latest. Therefore I have maintained my approach of dropping duplicates, but I would like to explore this better (maybe with the person that developed and used the scraper).

2.2. Airbnb Locations

The dataset “Mesh_Ids_Data_Itapema.csv” was also imported using Pandas, and was named **airbnb_locations**.

airbnb_locations is a Pandas Dataframe containing entries from "Mesh_Ids_Data_Itapema.csv" and it contains entries describing Airbnb listings **locations** scraped from the website.

2.2.1. Dropping duplicates

Some entries were duplicated due to the scraper executing on different days. They were sorted based on the day it was scrapped, and older entries were dropped.

2.2.2. Comparing lat/lon of *airbnb_locations* and *airbnb_listings*

Latitude and longitude from both tables (**airbnb_locations** and **airbnb_listings**) were compared and were not the same (although really close).

Since the case description is really specific about this - "*This dataset* [**airbnb_locations**] *contains the latitude and longitude for all available listings on airbnb for a given latitude and longitude square.* ***It is the most reliable data to infer the location of a listing***." - only the latitude and longitude from **airbnb_locations** were considered.

The latitude and longitude of **airbnb_listings** were replaced with the ones from **airbnb_locations**.

2.2.3. Encoding lat/long with Geohash

With the most reliable latitude and longitude, each entry location was encoded using geohashes of length 6 (1,22 km x 0,61 km rectangle) and length 7 (153 m x 153 m rectangle).

The geohash library used is "python-geohash"[2].

2.3. Airbnb price and availability

The dataset "Price_AV_Itapema-001.csv" was imported using Dask due to its increased size (4.6 GB) and was named **airbnb_price_av**. Spark could've also been used to work with this big dataset, but would require more work and many dependencies.

airbnb_price_av is a Dask Dataframe containing entries from "Price_AV_Itapema-001.csv" and contains entries describing Airbnb listings on a given date with features like availability of that listing at that day, price, minimum stay etc.

airbnb_price_av contains 43 Million entries.

Since each listing has a different price depending on week day and when it was scraped, the idea is to use this dataset to calculate an "average price" (mean, median, mode) of each listing.

All listings are in the same city, hence holiday prices will act on all listings equally, and clearly weekend prices will as well. That way, calculating an "average price" will suffice to compare prices among all listings.

[2] https://pypi.org/project/python-geohash/

2.3.1. Dropping *NaN* prices

Some entries had the feature ‘price’ a non-numeric value (NaN). Since it’s a small group (only 2986 entries), they were dropped.

2.3.2. Dropping unrealistic prices

*** This is one of the main challenges of the whole case. ***

The column ‘price’ is given and it is clearly a conversion from ‘price_string’. Therefore, if ‘price_string’ had a poorly formatted string, it created a misleading ‘price’.

Many entries had what it appears to be a “concatenated” entry in ‘price_string’, for example:

R$250280

Analyzing only one random listing with a poorly formatted string, many R$250,00 and R$280 appeared on ‘price_string’, therefore I concluded that the string should actually be, for example:

R$250

or

R$280

What probably happened is that the listing owner typed two values in sequence or there was a problem with the scraper.

Since it’s completely unrealistic for an Airbnb listing to cost R$250.280,00 for one night, I decided to remove all prices greater than R$100.000,00. Although not a perfect solution, it sufficed to complete the analysis.

Another approach that I have in mind but didn’t implement: each listing can vary its price a lot, especially during holidays and summer. Suppose that this price can increase 3 to 4 times between low and peak season. We could get the minimum price of each listing (corresponding to low season) and then drop all entries with prices greater than 3 or 4 times the minimum price. This way all the poorly formatted strings that appear to be concatenated would be eliminated.

2.3.3. Dropping NaN ‘av_for_checkin’

In a similar manner, column ‘av_for_checkin’ had some non-boolean values, and since there were not too many (only 37.738 entries), they were dropped.

2.3.4. Calculate average price

Calculating the “mean” would result in low quality values, since some entries would still have poorly formatted strings (which in turn resulted in a wrong ‘price’ value).

Dask doesn't support calculating the median yet. Calculating the median using parallel computing can be a hard task, although it could have resulted in better averages.

Therefore the average function used to calculate each listing price is the "mode".

After calculating the 'mode price' of each listing, its value was included on **airbnb_listings**.

2.3.5. Calculate availability rate

Each listing was grouped together using 'airbnb_listing_id' and the method 'value_counts' was used to count how many days were 'Available' or 'Not available'.

Availability rate was calculated, for each listing, as following:

$$Availability\ rate = \frac{Total\ available\ days}{Total\ available\ days + Total\ unavailable\ days}$$

After calculating the 'availability rate' of each listing, its value was included on **airbnb_listings**.

2.4. Most sought-after listings

Ideally an investor wants his/her property to be rented most of the time with the greatest price possible. There's no official vacancy rate in the dataset, but they it can be estimated in two forms[3]:

1. Availability rate: a low availability indicates that the property is usually in high demand and rented out (remember: less availability is better!)
2. Number of reviews: more reviews indicates that the property is rented more often, otherwise guests wouldn't be able to review it.

A small caveat: a low availability does not necessarily indicate that the property was rented most of the time. Maybe the owner decided that he/she didn't want to rent it at that time but still wanted the listing posted on Airbnb's platform.

Therefore both metrics will be utilized, but Number of reviews will be favored over Availability rate.

After analyzing **airbnb_listings**, I have decided to set a threshold of at least 50 reviews to select only the top booked Airbnb listings (it was stored in a variable called **top_booked_airbnb_listings**).

Similarly, I have decided to set a threshold of at most 10% availability rate on **airbnb_listings** to select the least available Airbnb listings (it was stored in a variable called **least_available_airbnb_listings**).

[3] Further analysis needs to be made comparing 'Availability rate' and 'Number of reviews' to assess which feature can better represent the vacancy of a listing.

2.5. Selecting only apartments

When analyzing the **top_booked_airbnb_listings** and **least_available_airbnb_listings** it became clear that the most popular listing type is the apartment (data that led me to this conclusion will be presented in Data Analysis).

Therefore another division of the data was made, selecting only apartments from **top_booked_airbnb_listings** and creating **top_booked_airbnb_apartments.** The division could have been made on **least_available_airbnb_listings** but it wasn't due to the caveat when creating it.

2.6. VivaReal listings

The dataset "VivaReal_Itapema.csv" was imported using Pandas and was named **vivareal_listings**.

vivareal_listings is a Pandas Dataframe containing entries from "VivaReal_Itapemacsv" and contains entries describing VivaReal listings **characteristics** scraped from the website.

The main idea was to use this dataset to calculate the square meter price of lots and apartments in a given region (the one defined in Question 4).

This dataset doesn't contain latitude and longitude, so it's not possible to encode each listing location into a geohash. The only locations available are state, city, neighborhood and, rarely, street.

Cleaning on this dataset consisted mainly of removing duplicate entries, removing accents from location strings and dropping entries with an unrealistic square meter price (only 1 entry was dropped).

Therefore I have decided to simply group listing by neighborhood and then assess how many properties were being sold in each neighborhood.

I have also divided **vivareal_listings** into two dataframes:

1. Lots: only lots being sold, named **vivareal_lots_selling**
2. Apartments: only apartments being sold, named **vivareal_apartments_selling**

Finally, both **vivareal_lots_selling** and **vivareal_apartments_selling** were grouped by neighborhood to create **vivareal_lots_nbh** and **vivareal_apartments_nbh**, respectively. For each group, the count of how many properties are in the neighborhood and the mean price/m² were calculated.

Transforming the neighborhood to a corresponding geohash would greatly improve this analysis. Maybe it could be done using Google Maps Api for example. Grouping by the neighborhood is a first step, but it would be great to capture the latitude and longitude of these listings.

It would also be really important to group listings that have sea view or not.

3. Real estate development cost and ROI estimate

In order to answer question 5, it’s

Calculations of Return on Investment relies on an Excel spreadsheet developed by me and improved over the years. This spreadsheet is tested and approved in the real world and was used to correctly estimate costs of a 12 apartment building built between 2019 and 2021.

The spreadsheet was updated to use data from Santa Catarina, Itapema (taxes, zoning regulations, cost of construction).

The cost of construction is from “Sinduscon de Balneário Camboriú”.

The cost of lots and apartments in a given region were manually acquired from VivaReal. The reason for that is discussed in section “Data Analysis - Question 4”.

4. Data Analysis

The data analysis portion of the case used bar plots, count plots and histograms to indicate characteristics of each different subset of the data.

Folium was used to create maps with the geohashes colored, indicating a given feature magnitude in terms of color intensity.

In the following subsections each plot will be presented with the corresponding dataframe and column used to generate it.

4.1. Question 1: What is the best property profile to invest in the city?

The best property profile to invest in Itapema is a:

- Apartment
- Bedrooms: 2 or 3
- Bathrooms: 1 or 2
- Beds: 2 to 4 double beds (to accommodate up to 6 guests)
- Location: see question 2
- Others characteristics: see question 3

Property profile 1 (preferred):

- **Apartment**
- **Bedrooms: 2**
- **Bathrooms: 1**
- **Beds: 2 double beds**
- **Location: see question 2**
- **Others characteristics: see question 3**

Property profile 2:

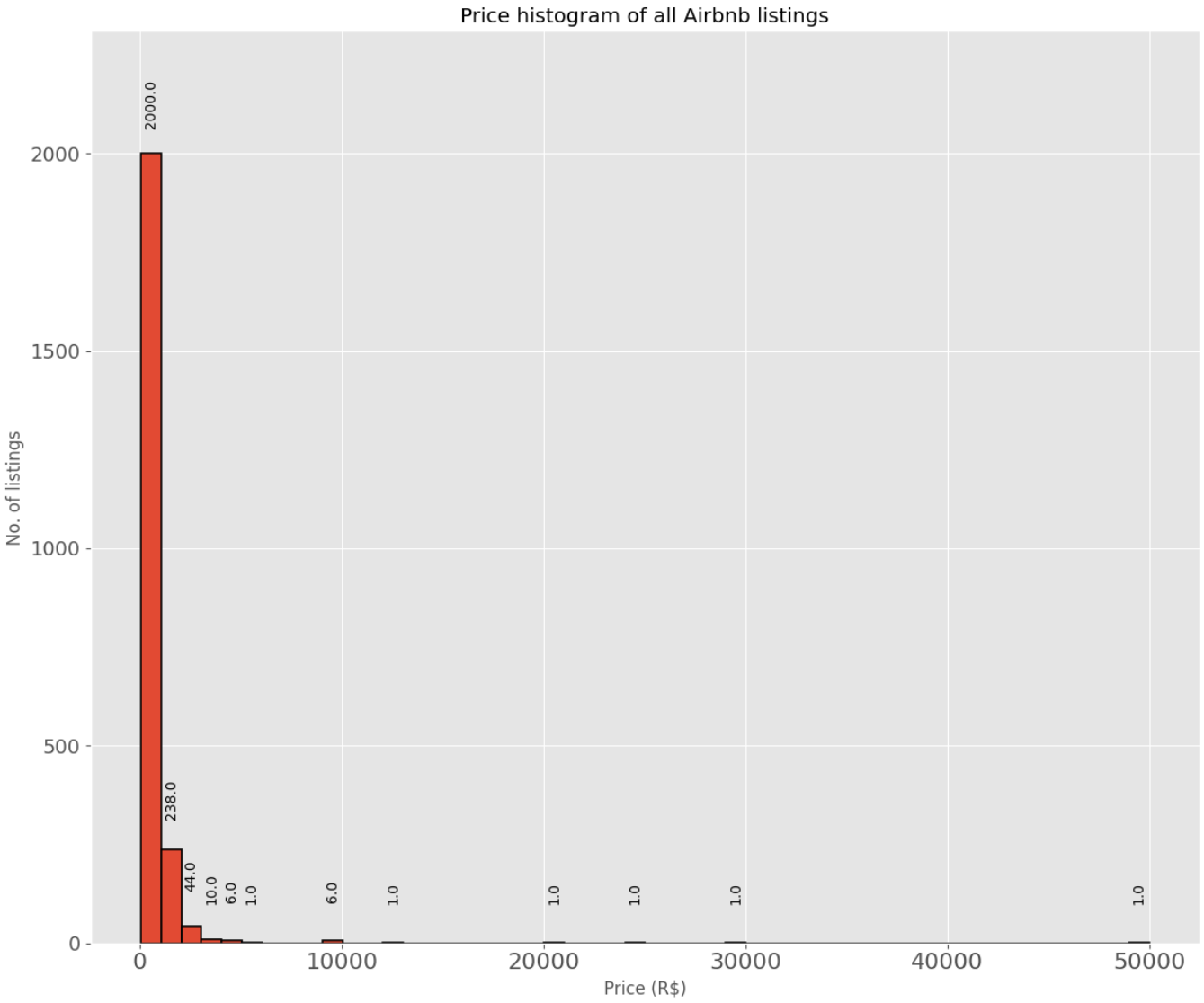
- Apartment
- Bedrooms: 3
- Bathrooms: 2
- Beds: 3 double beds + 1 single bed

In terms of investment, a smaller property usually generates a higher Return on Investment, so **the best property profile would be Property profile 1**.

Below are the analyses made to come up to this conclusion.

4.1.1. Price histogram among all listings on Airbnb

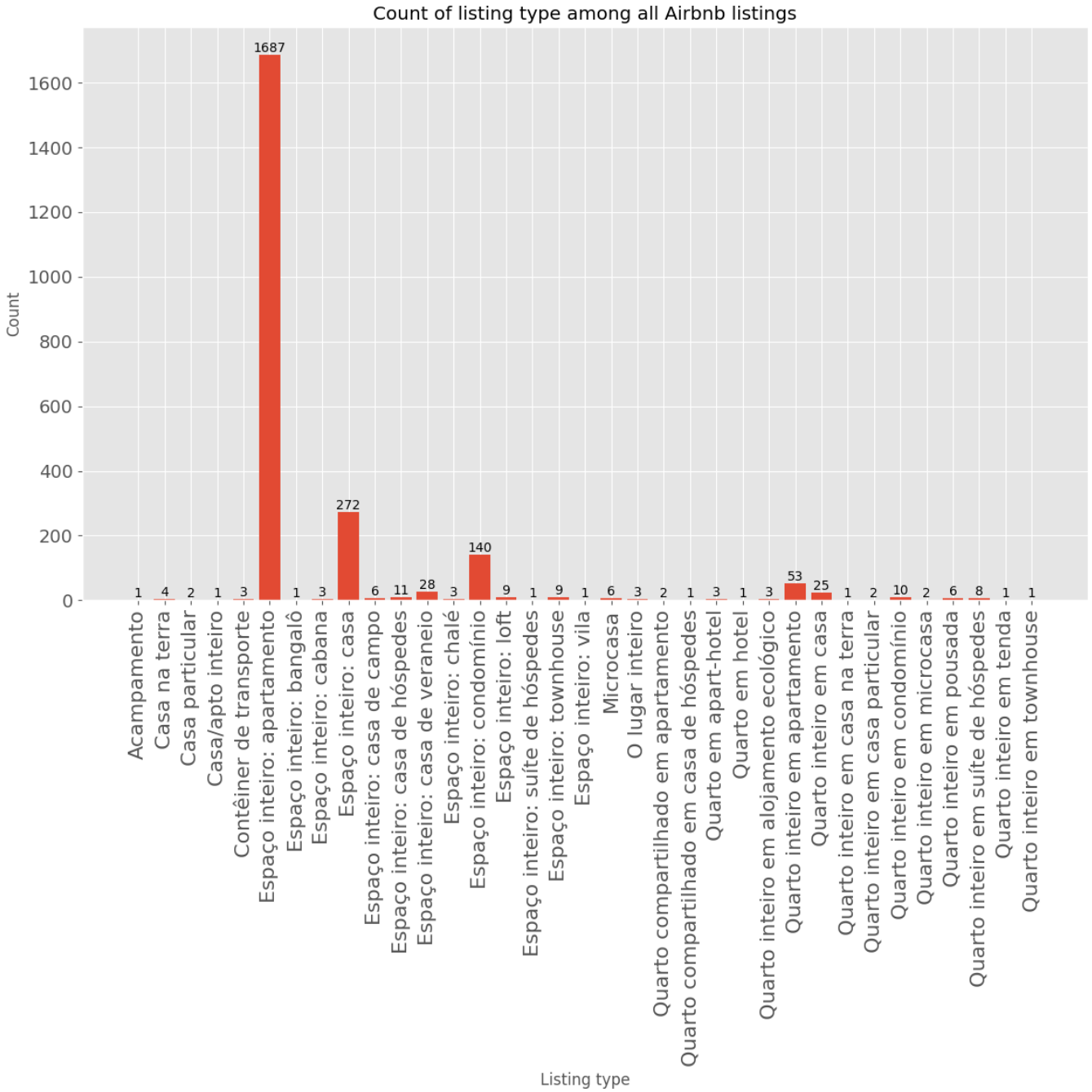
- dataframe = airbnb_listings
- column = mode_price



Naturally most listings are under R$5.000,00/night. This plot makes it clear the problem stated in section “Airbnb price and availability” regarding poorly formatted strings. Many poorly formatted entries were dropped, but there are still entries present in the dataframe.

4.1.2. Listing types among all listings on Airbnb

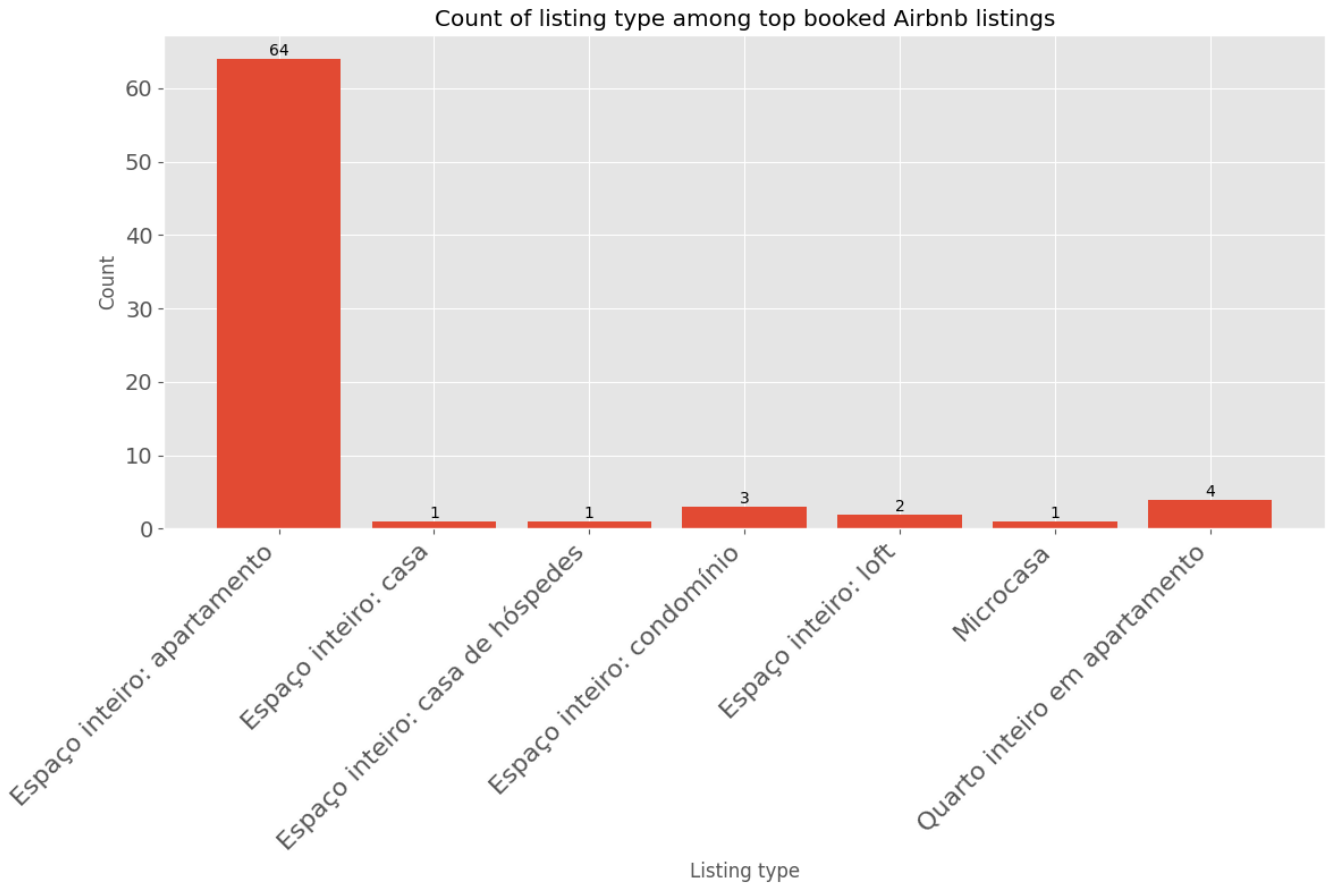
- dataframe = airbnb_listings
- column = listing_type



The vast majority of listings on Airbnb are apartments. Houses are also present, but nearly 7 times less. This is one of the reasons that only apartments will be analyzed further.

4.1.3. Listing types among top booked listings on Airbnb

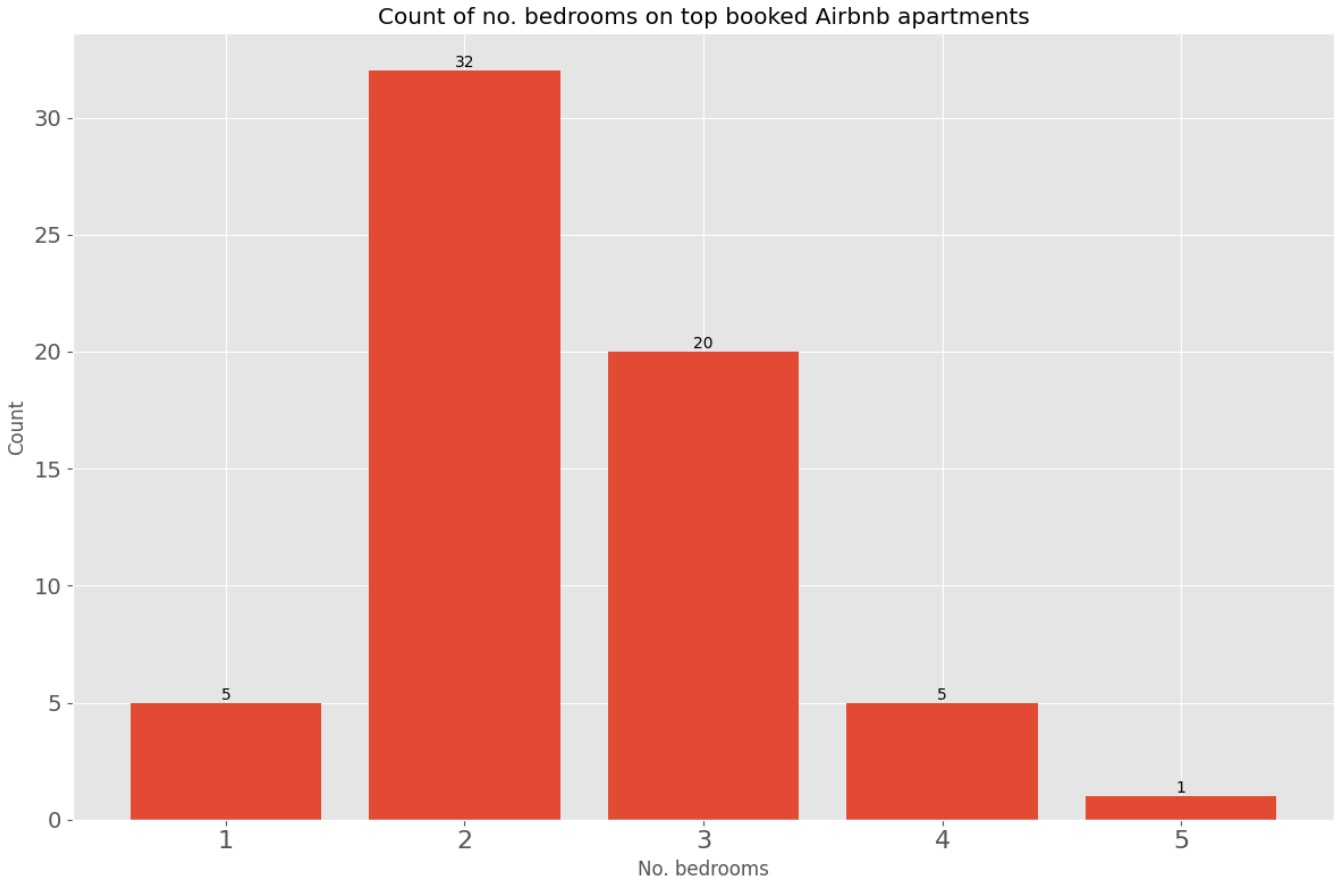
- dataframe = top_booked_airbnb_listings
- column = listing_type



Among the top booked Airbnb listings, it's clear that most are apartments. **Therefore the analysis will focus on this type of real estate.**

4.1.4. Number of bedrooms among top booked Airbnb apartments

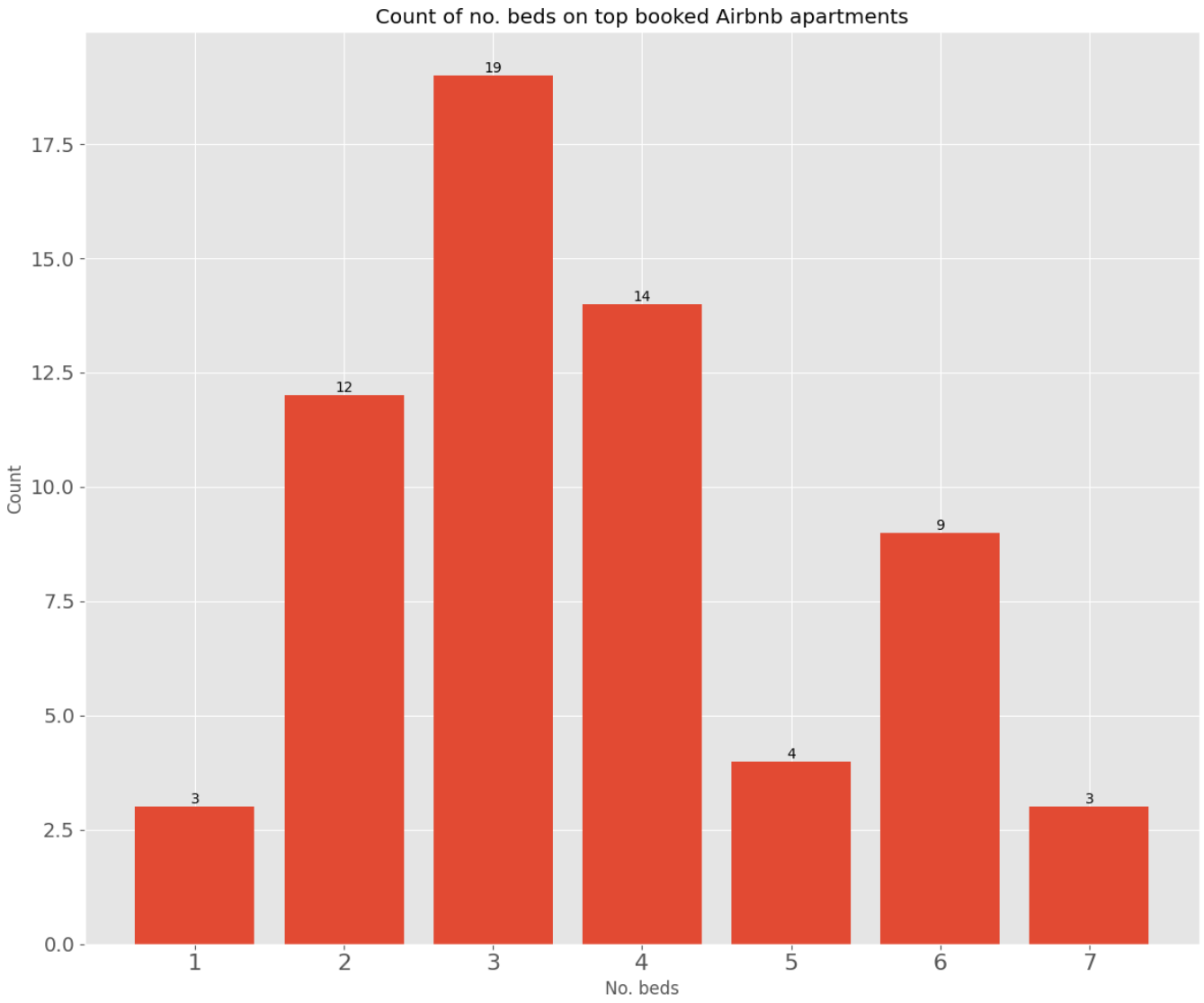
- dataframe = top_booked_airbnb_apartments
- column = number_of_bedrooms



Among top booked Airbnb apartments, **the most usual number of bedrooms is two**. Three bedrooms are also really relevant.

4.1.5. Number of beds among top booked Airbnb apartments

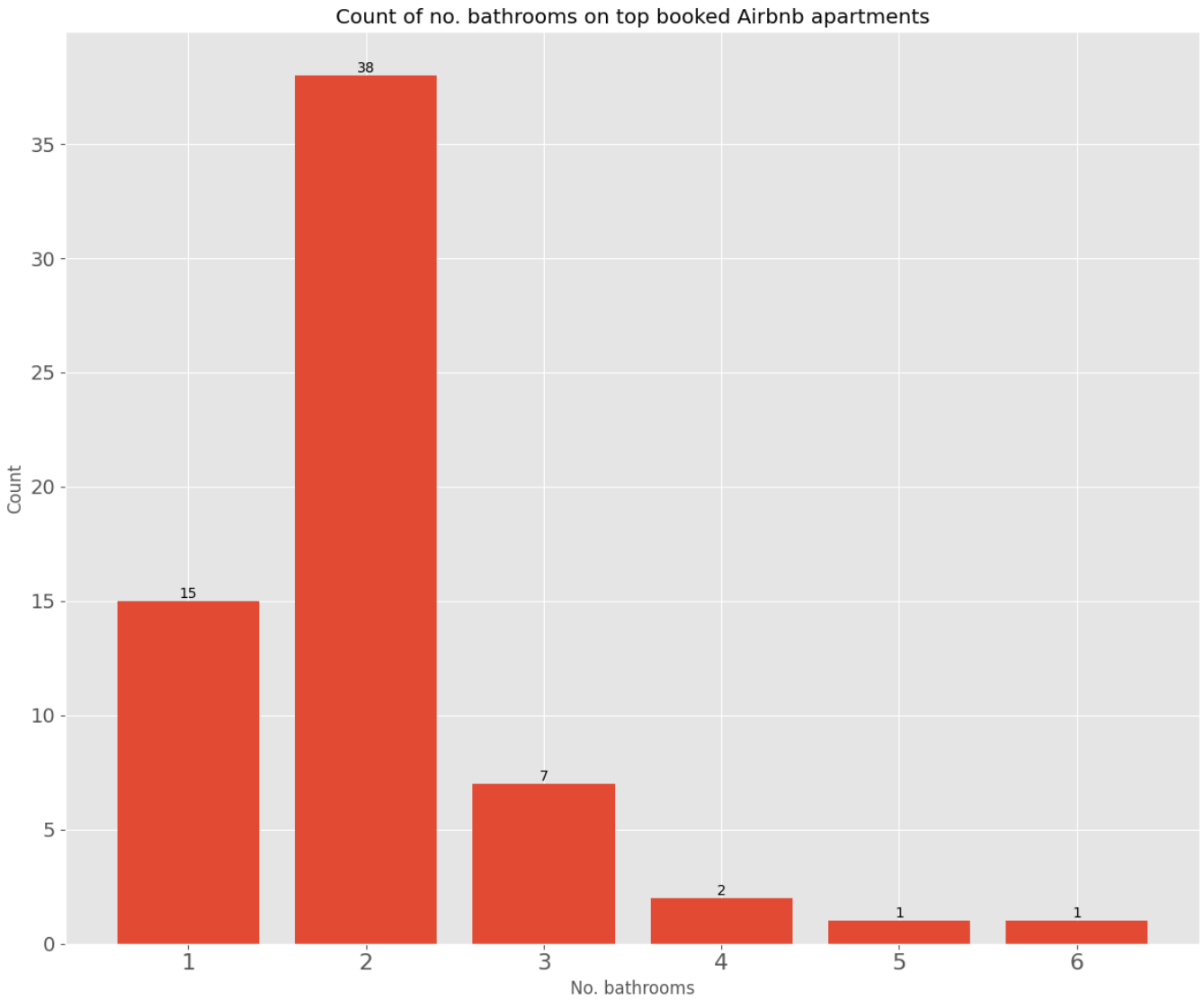
- dataframe = top_booked_airbnb_apartments
- column = number_of_beds



Top booked Airbnb apartments usually have between 2 and 4 beds, but it's not specified if they are single or double beds. It will become more clear.

4.1.6. Number of bathrooms among top booked Airbnb apartments

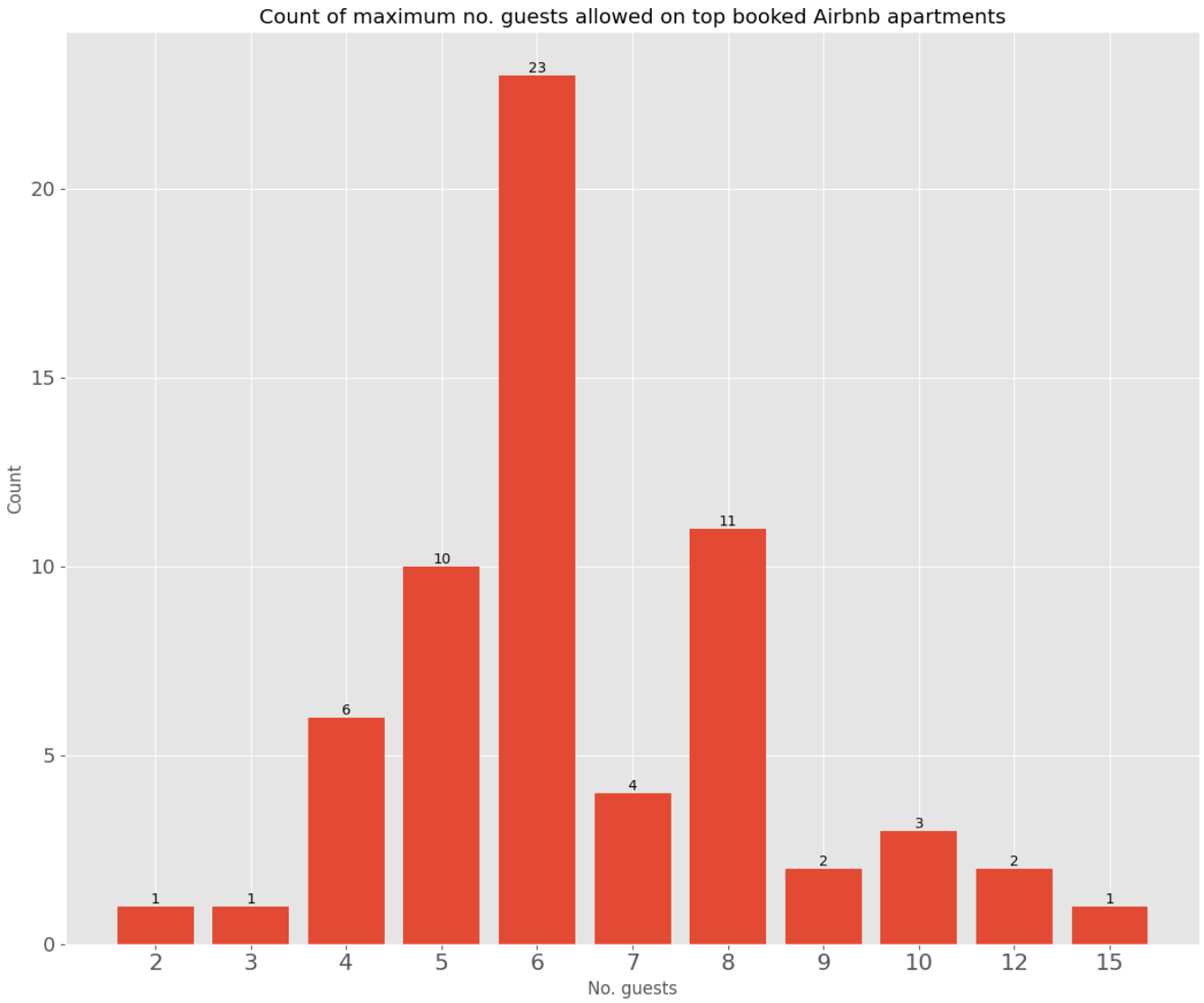
- dataframe = top_booked_airbnb_apartments
- column = number_of_bathrooms



One or two bathrooms are the most usual on top booked Airbnb apartments.

4.1.7. Maximum no. guests among top booked Airbnb apartments

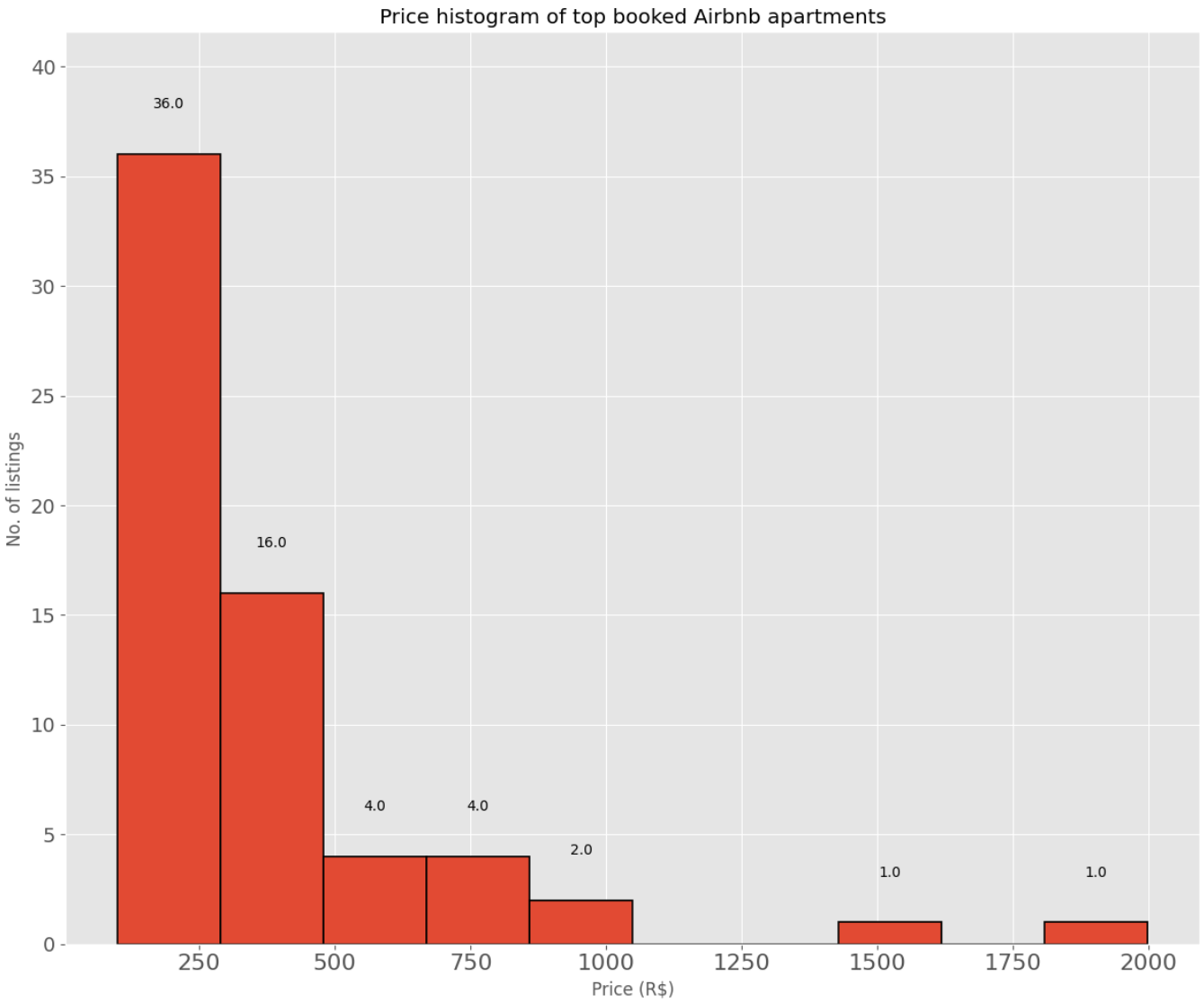
- dataframe = top_booked_airbnb_apartments
- column = number_of_bathrooms



Top booked Airbnb apartments **usually accommodate up to 6 guests**, but some can accommodate up to 8! Combined with the number of beds, it's possible to conclude that **most beds are double beds**.

4.1.8. Price histogram among top booked Airbnb apartments

- dataframe = top_booked_airbnb_apartments
- column = mode_price



Top booked Airbnb apartments didn't include any entry with a poorly formatted string. If the poorly formatted strings are caused by the listing owner (and not the scraper) this is a good indicator (and obvious) that a good quality listing increases its chance of being rented (idea: calculate by how much if it's relevant to show this information to Seazone's hosts).

4.2. Question 2: Which is the best location in the city in terms of revenue?

Canto da Praia, Meia Praia, Morretes (near the sea) and Centro are all neighborhoods with the highest price/night.

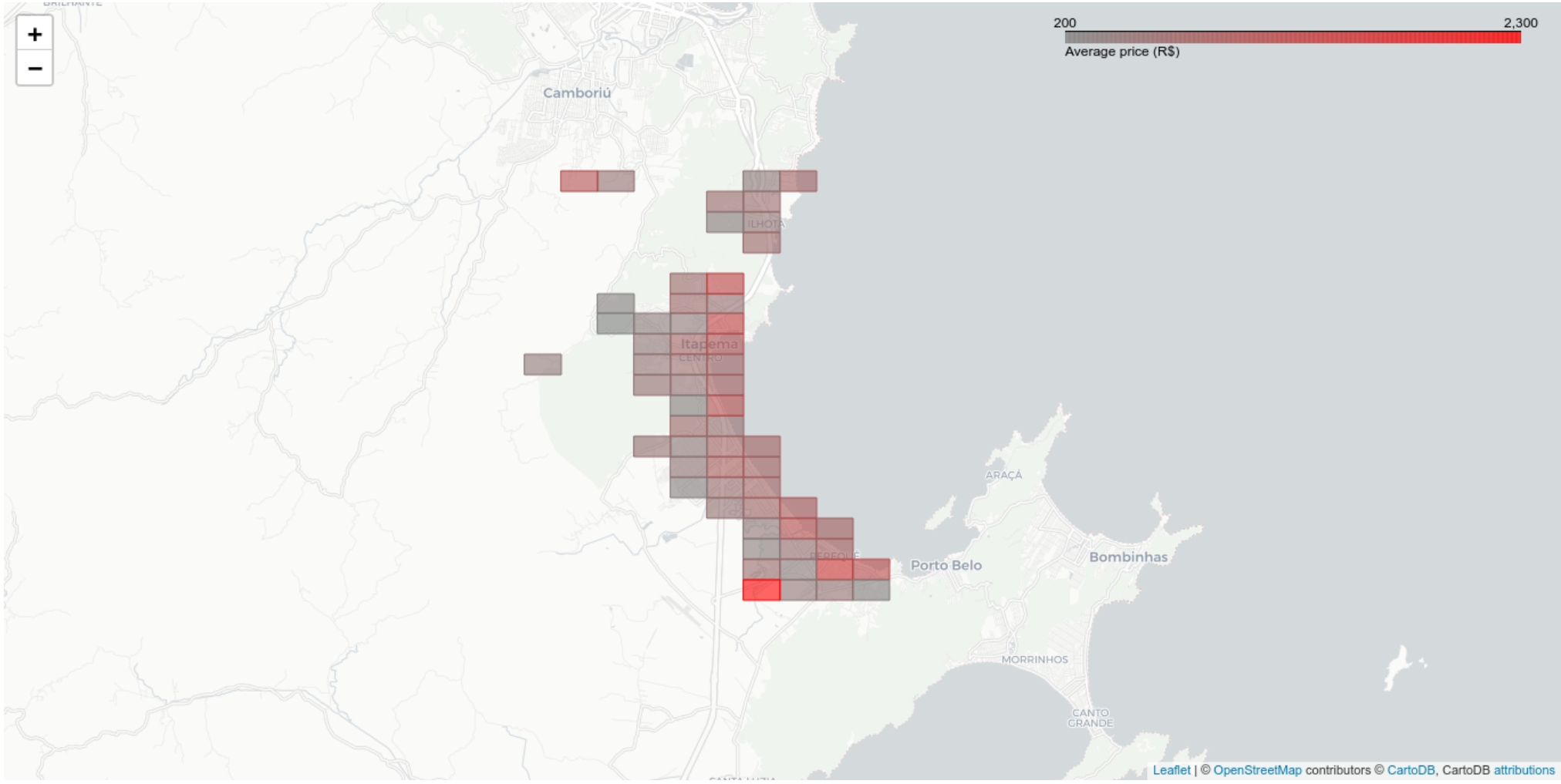
The best locations in terms of revenue though are Canto da Praia and Meia Praia, followed by Centro and Morretes (near the sea).

Below are the analyses made to come up to this conclusion.

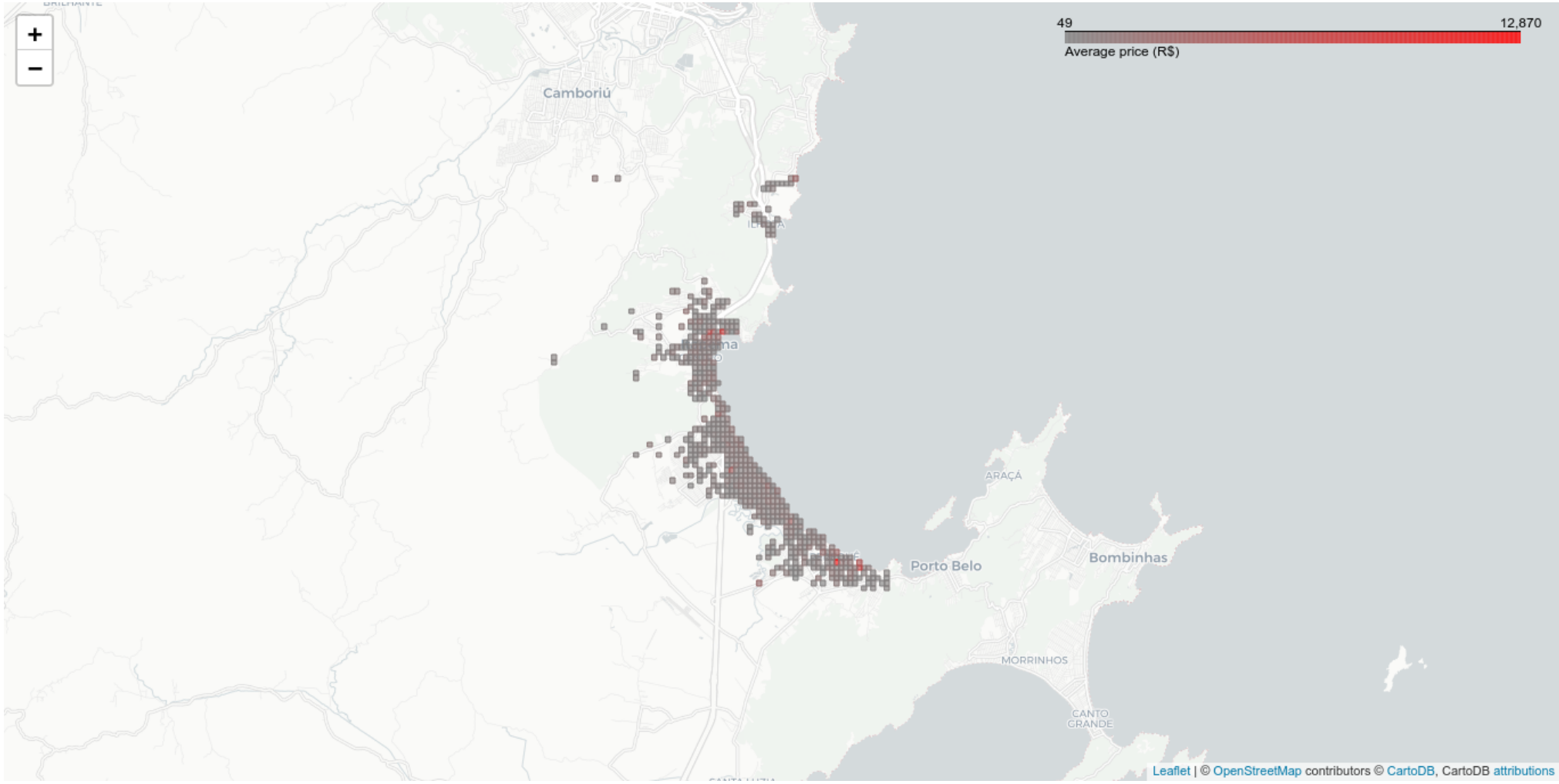
4.2.1. Map plot of all Airbnb listings with respect to mean price

All map analyses were performed using Google Maps to get the name of the neighborhood.

- dataframe = airbnb_listings_geohash_p6 and airbnb_listings_geohash_p7



Naturally, regions close to the beach are more expensive. There's an outlier on Porto Belo that is actually far from the sea.

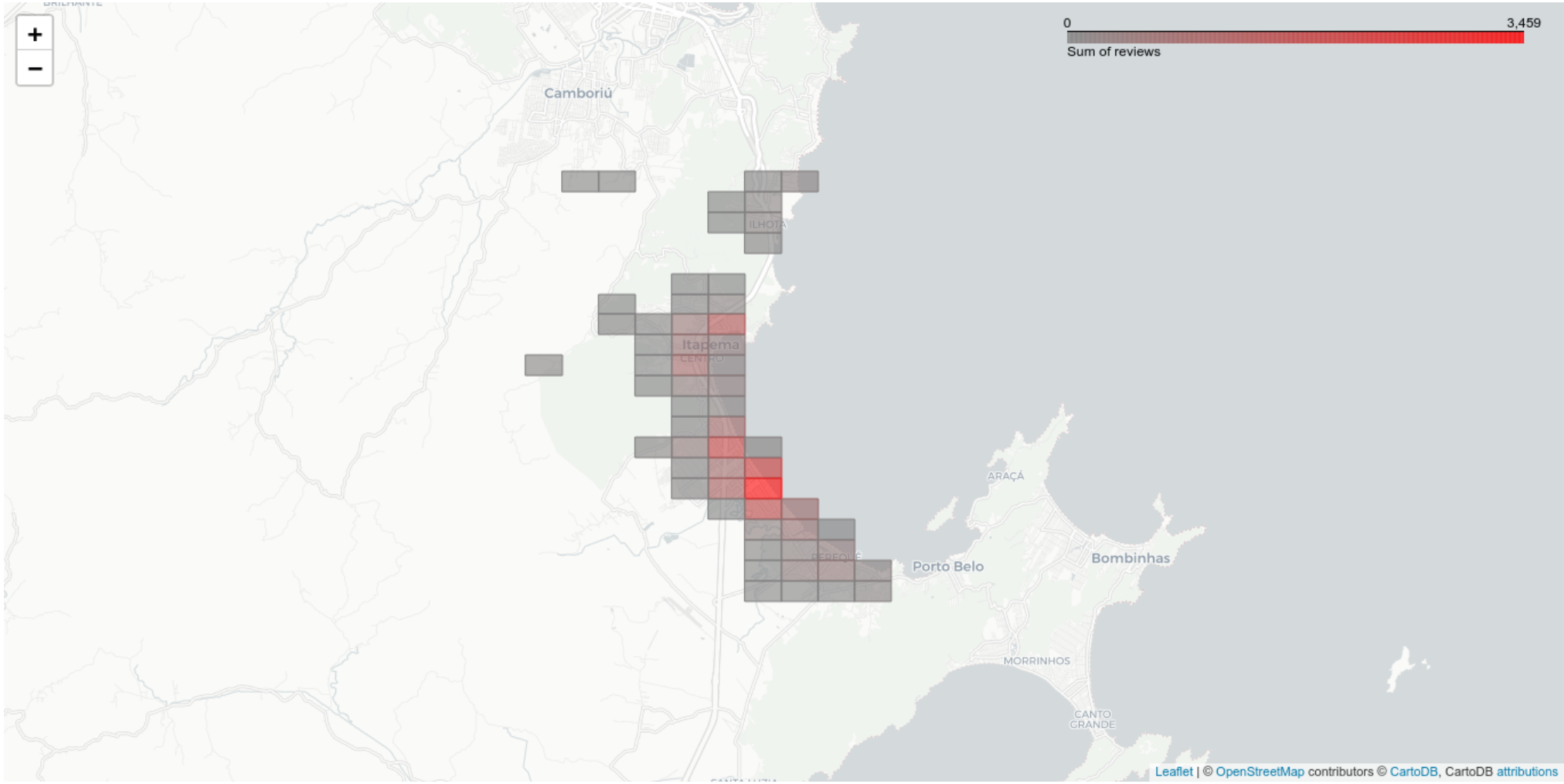


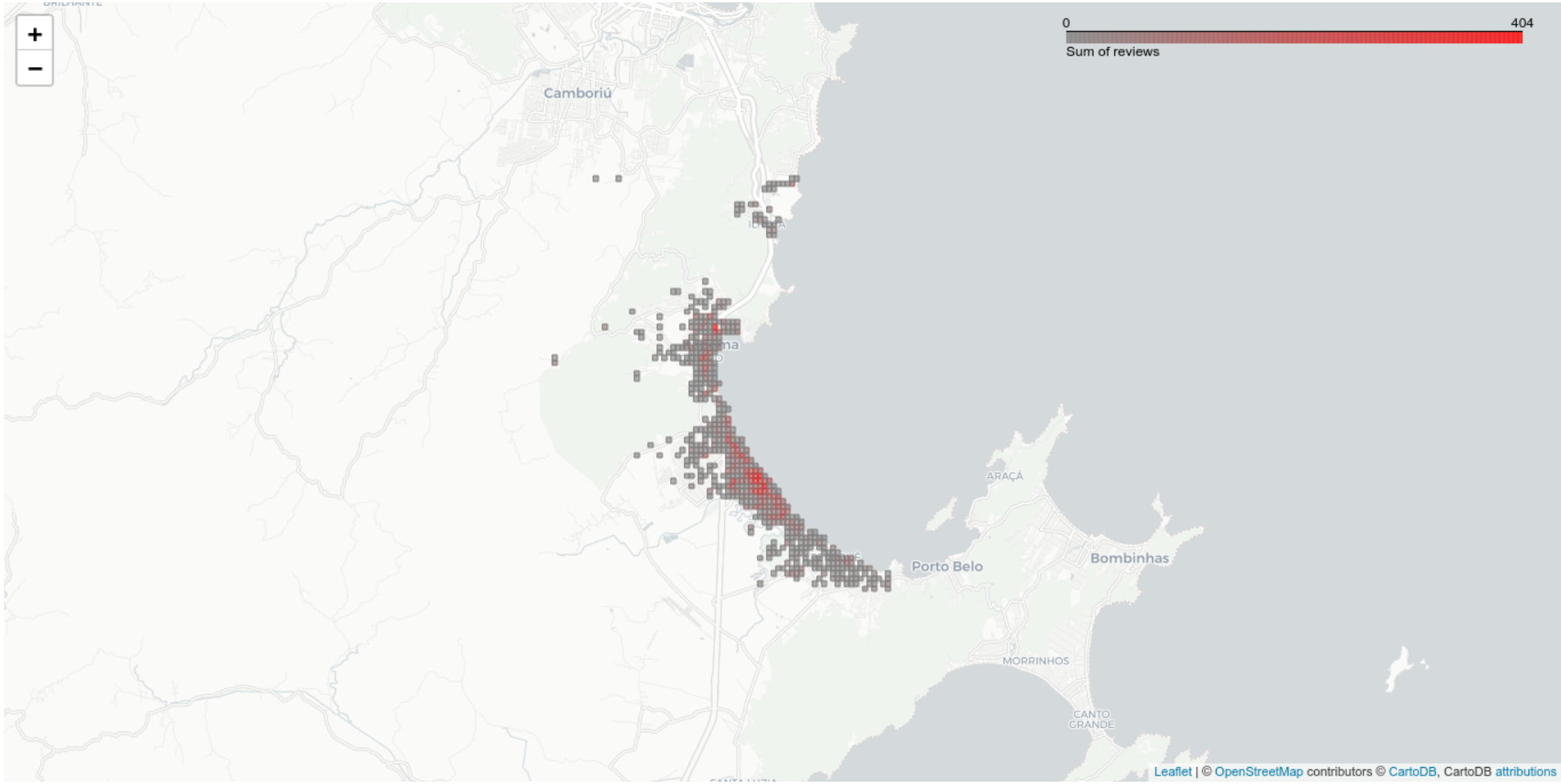
When dividing the regions using a more precise geohash, the differences are spread out and outliers make most regions appear "dim".

But it's clear that the region Canto da Praia and Porto Belo have the highest average price (but we will analyze this again using only top booked Airbnb apartments).

4.2.2. Map plot of all Airbnb listings with respect to no. reviews

- dataframe = airbnb_listings_geohash_p6 and airbnb_listings_geohash_p7



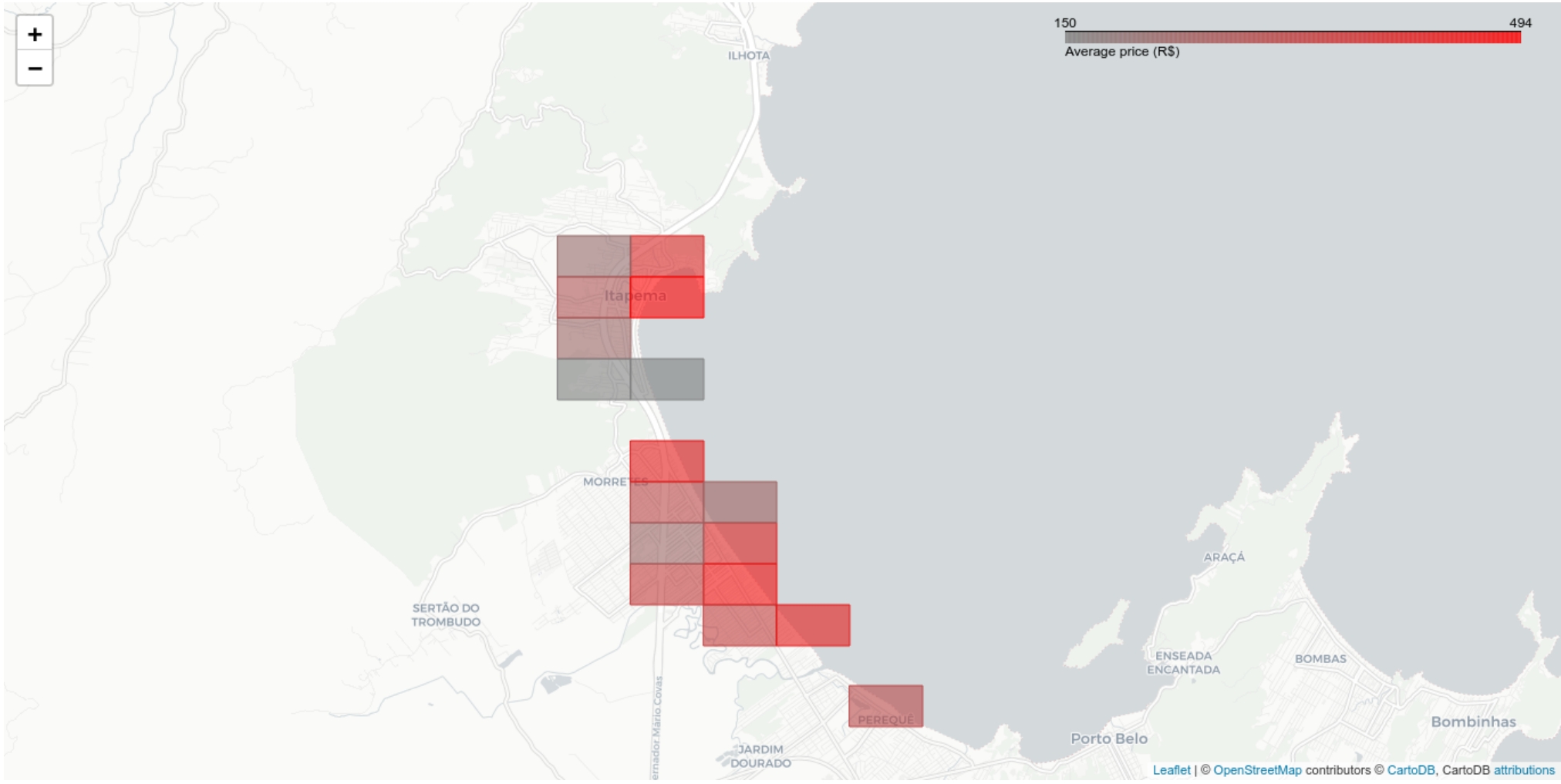


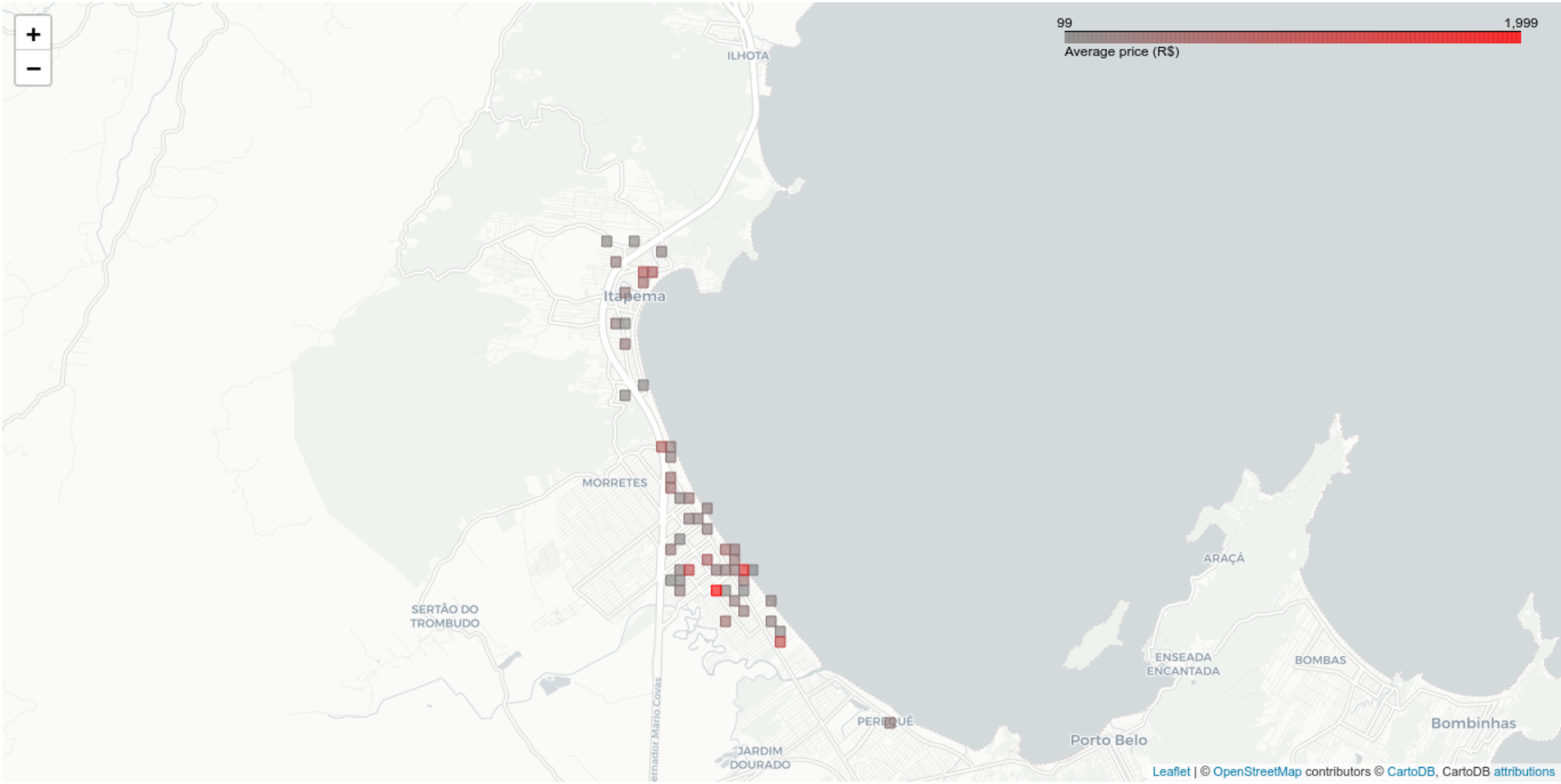
Meia Praia, Morretes, Canto da Praia and Centro all have a high number of reviews, as long as it's close to the sea, indicating that they are preferred by guests!

4.2.3. Map plot of top booked Airbnb apartments with respect to price

Now we focus our attention on only top booked Airbnb apartments.

- dataframe = top_booked_airbnb_apartments_geohash_p6 and top_booked_airbnb_apartments_geohash_p6

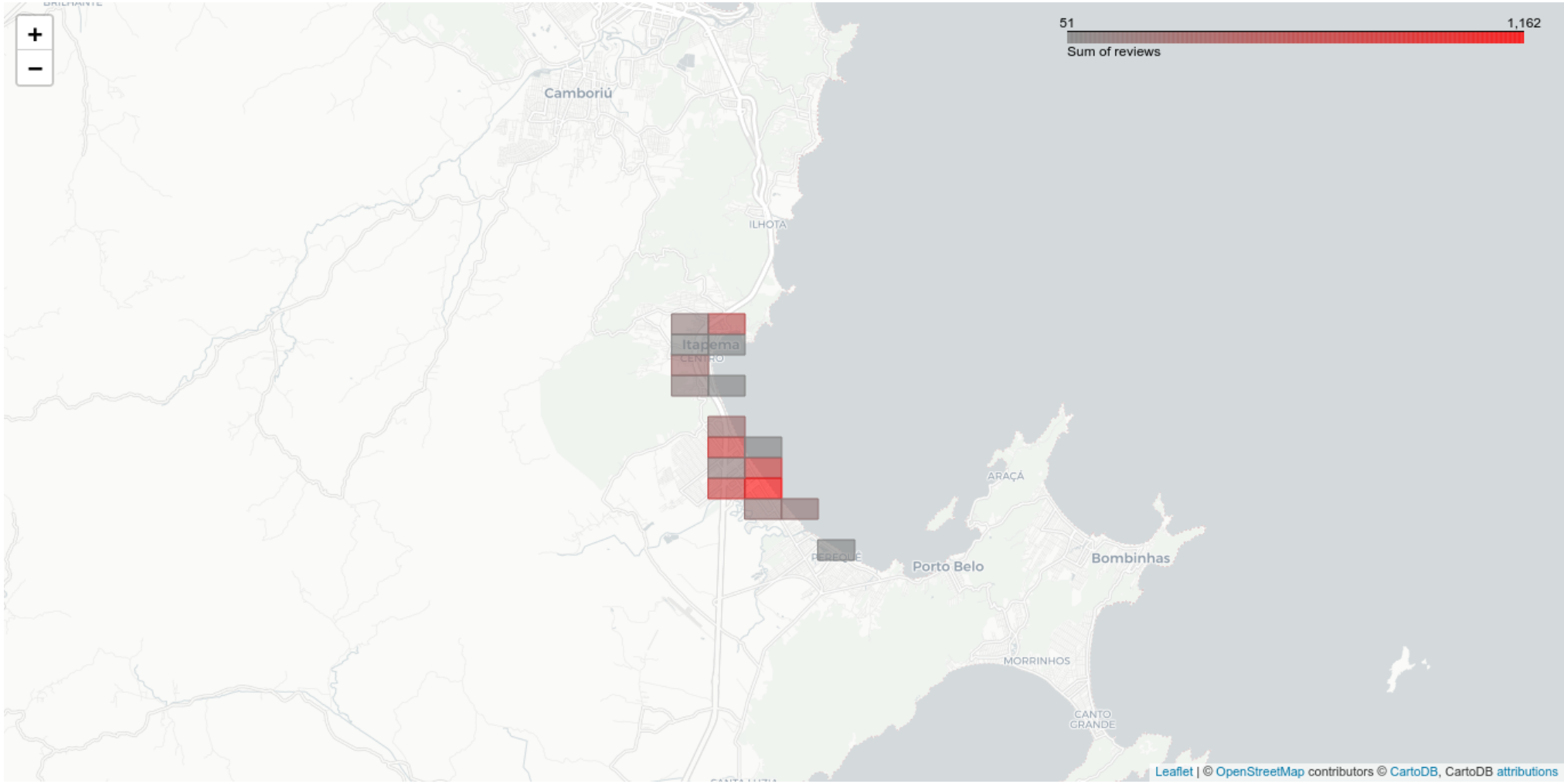


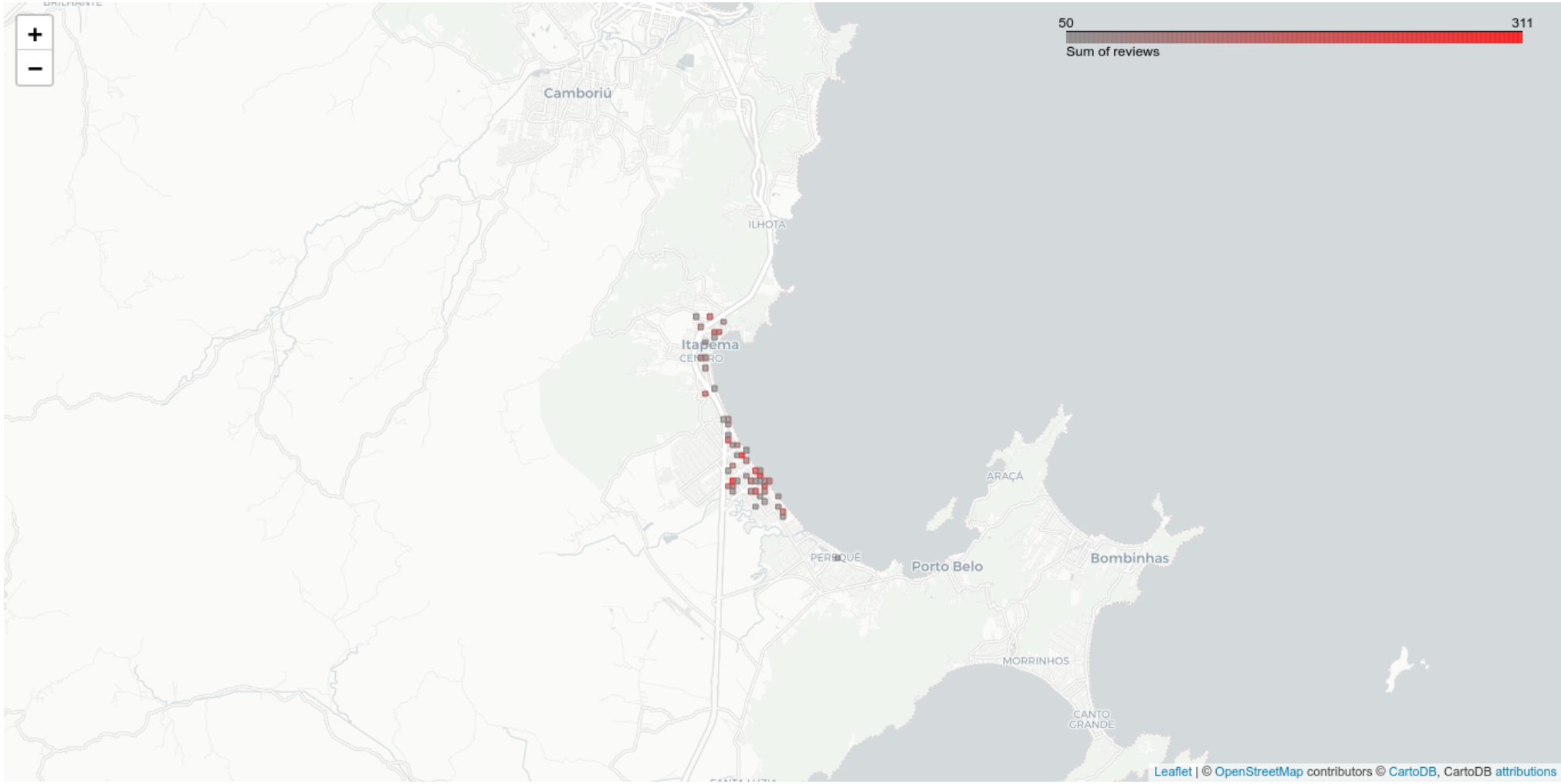


The trend becomes clear now: Meia Praia, Morretes, Centro, Canto da Praia are all regions that the top booked Airbnb apartments have the highest price/night!

4.2.4. Map plot of top booked Airbnb apartments with respect to no. reviews

- dataframe = top_booked_airbnb_apartments_geohash_p6 and top_booked_airbnb_apartments_geohash_p6





Among the top booked Airbnb apartments, the ones with the highest number of reviews (indicating that they are preferred) are in **Meia Praia, Morretes and Canto da Praia.**

4.3. Question 3: What are the characteristics and reasons for the best revenues in the city?

As already defined, part of the characteristics for the best revenues are all of those listed on Question 1, **Property profile 1**.

But there are many other characteristics, for example, having a pool, wifi, TV and so on. All of those other characteristics were extracted from Airbnb listings and one-hot-encoded. They are presented below.

To summarize, the main characteristics that top booked airbnb apartments (with the highest revenues) have are:

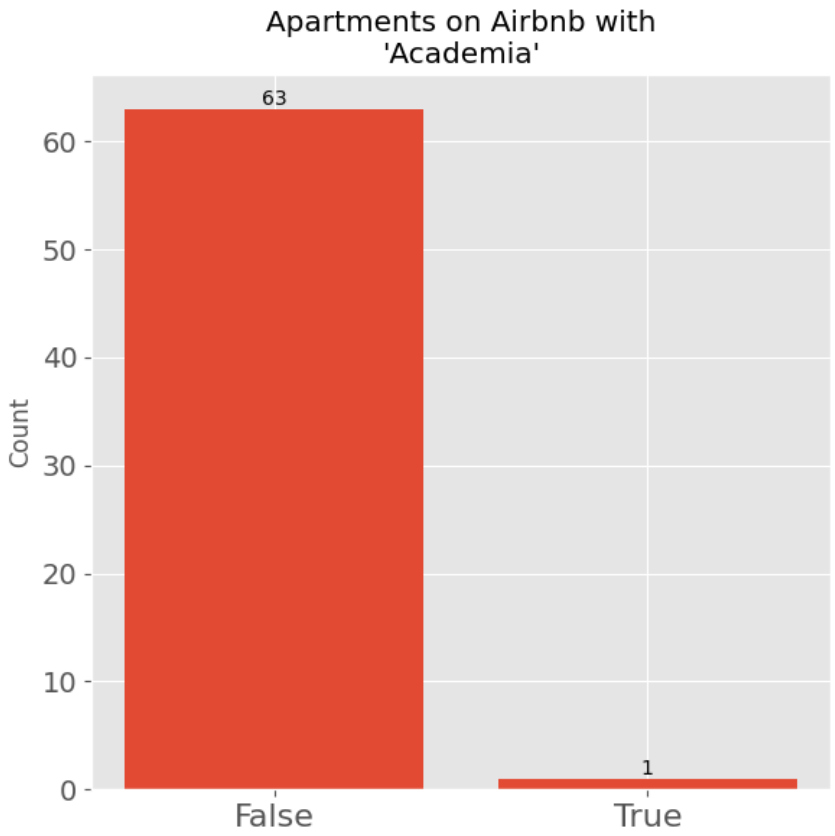
- **Profile**
  - **Apartment**
  - **Bedrooms: 2 or 3**
  - **Bathrooms: 1 or 2**
  - **Beds: 2 to 4 double beds (to accommodate up to 6 guests)**
- **Location: Canto da Praia, Meia Praia, Morretes and Centro**
- **Characteristics:**
  - **No gym**
  - **Pets being allowed is unimportant**
  - **Has air conditioning**
  - **Has at least one parking spot**
  - **Kitchen with:**
    - **Stove**
    - **Microwave**
    - **Refrigerator**
    - **Kitchen essentials**
  - **Having beach essentials is unimportant**
  - **Has a washing machine**
  - **No pool**
  - **Bedding is unimportant**
  - **Has a TV**
  - **Has wifi**
  - **Sea view is unimportant**
  - **Since it's a short term stay, furnitures are always present**

4.3.1. Encoded columns of top booked Airbnb apartments

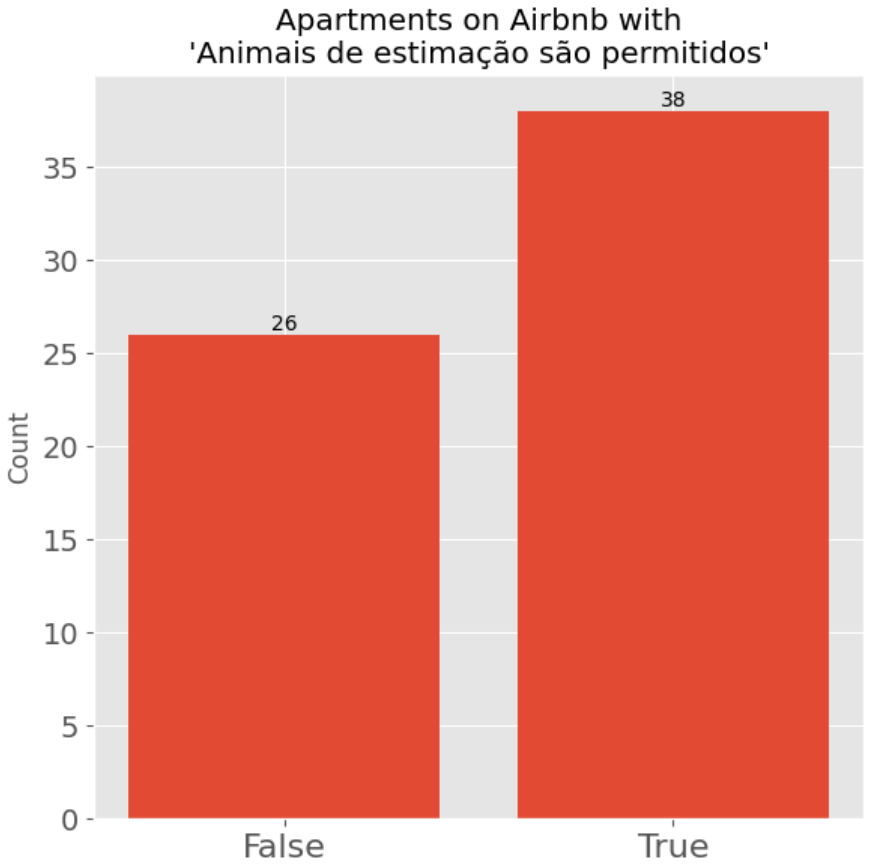
- dataframe = top_booked_airbnb_apartments

- column = all of the one-hot-encoding columns

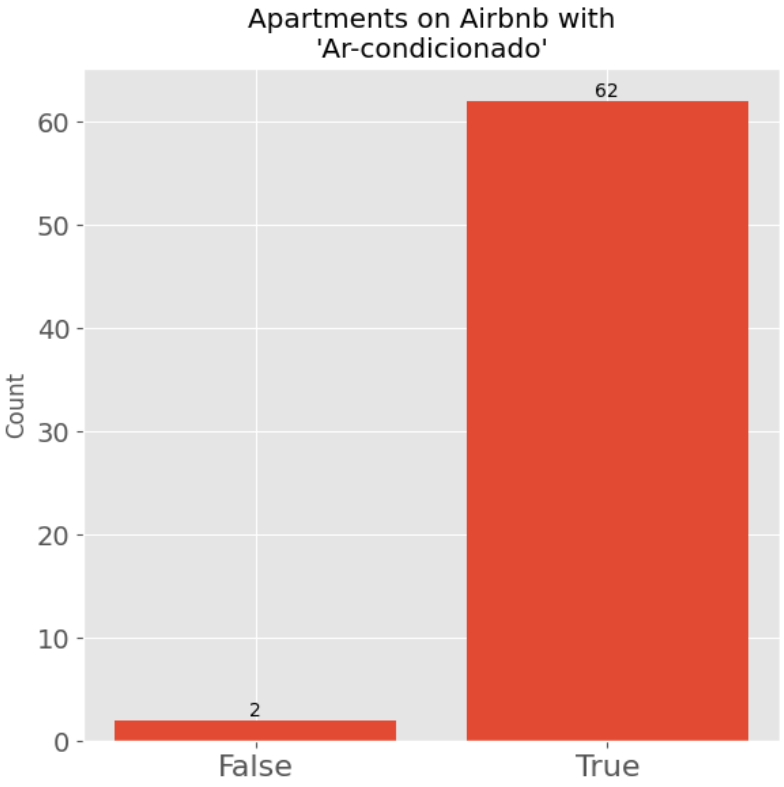
There are 130 columns that were encoded using one-hot-encoding, so I will just display the features that I find more relevant to characterize a listing.



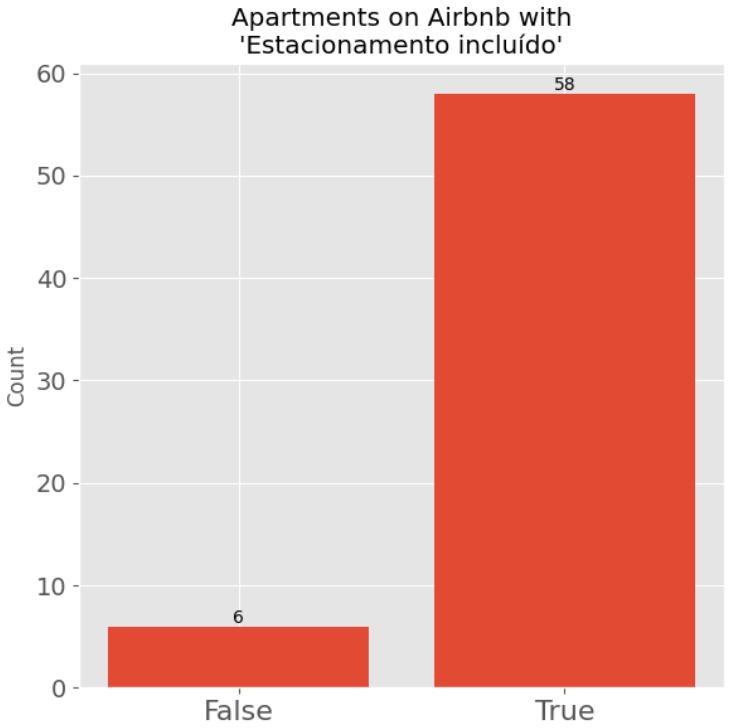
Having a gym is not relevant.



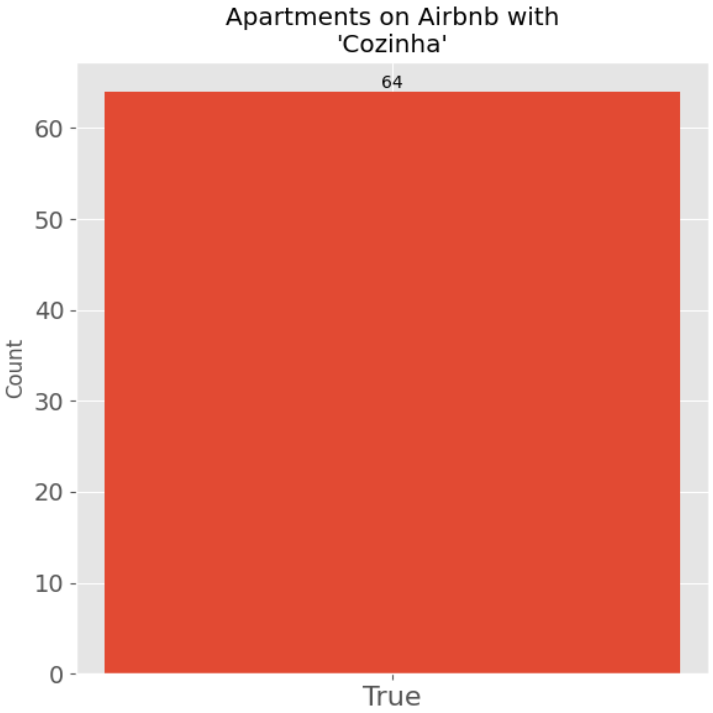
Pets are usually allowed, but it’s not decisive.



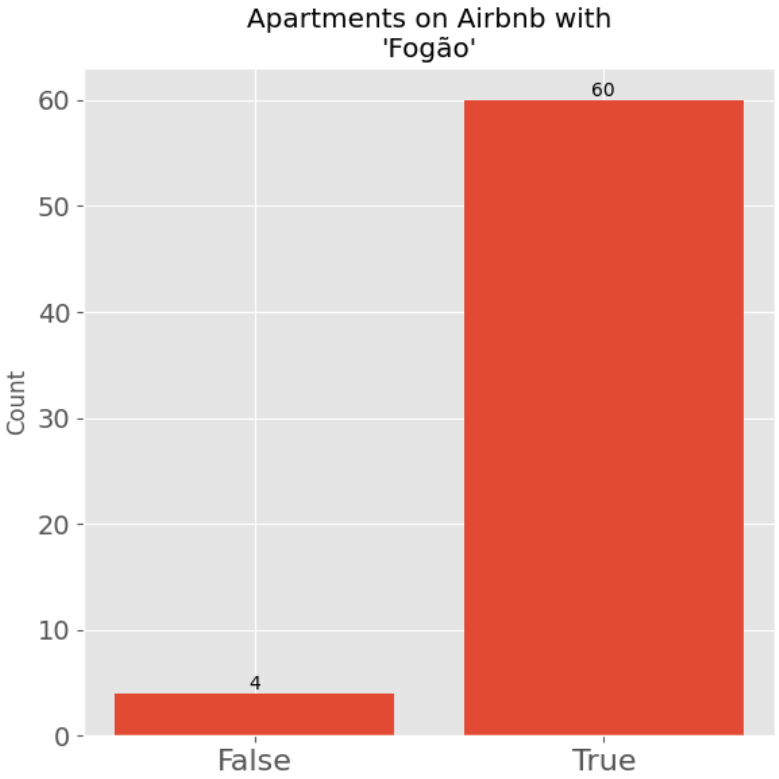
Having an air conditioner is almost a rule on top booked Airbnb apartments.



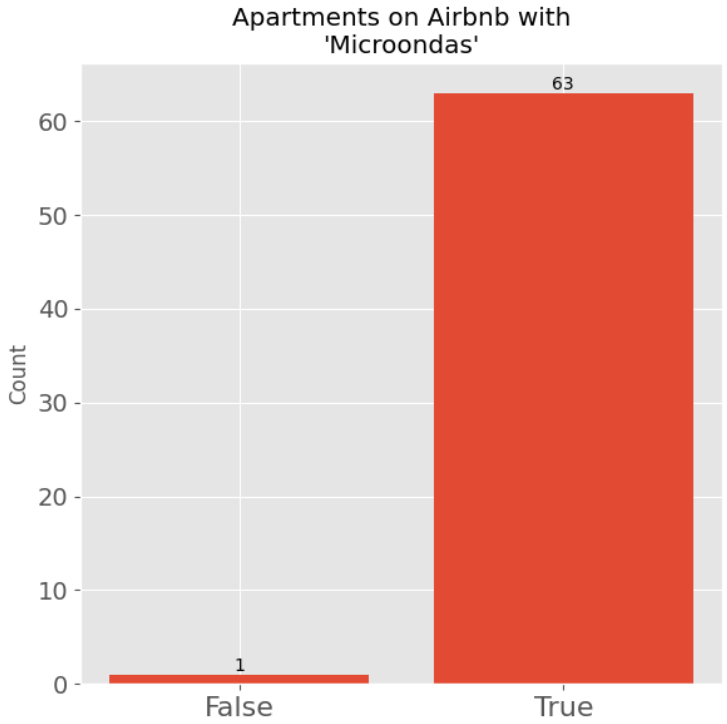
Having at least one parking space is almost a rule on top booked Airbnb apartments.



All top booked Airbnb apartments have a kitchen.

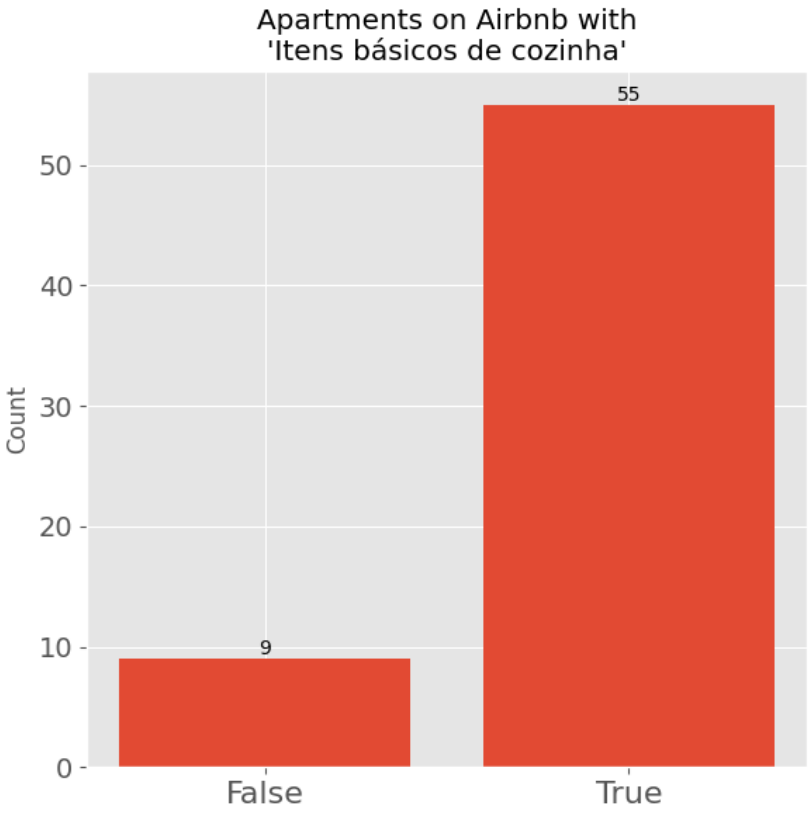


Having a stove in the kitchen is almost a rule on top booked Airbnb apartments.

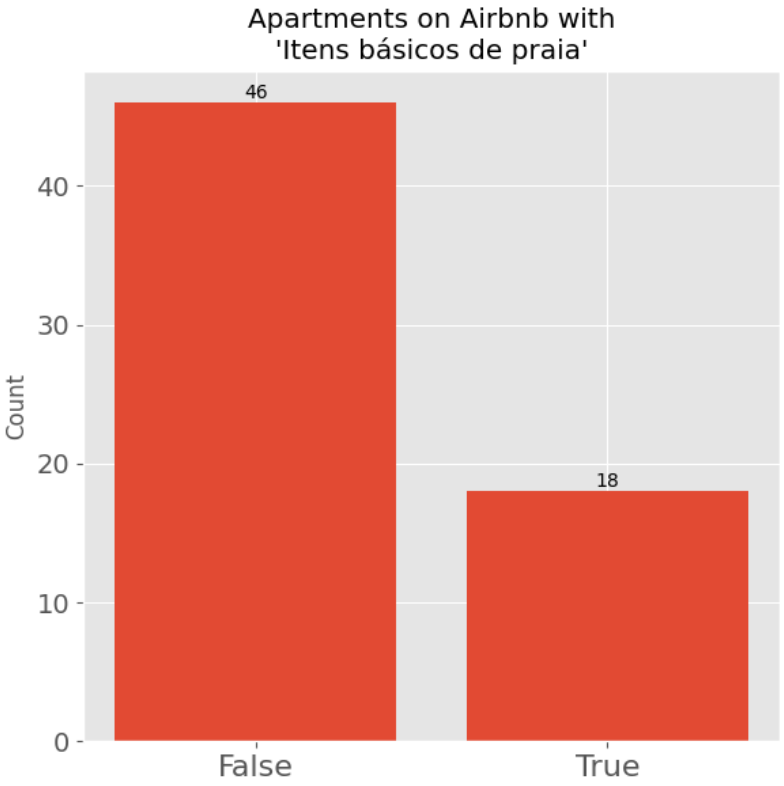


Having a microwave in the kitchen is almost a rule on top booked Airbnb apartments.

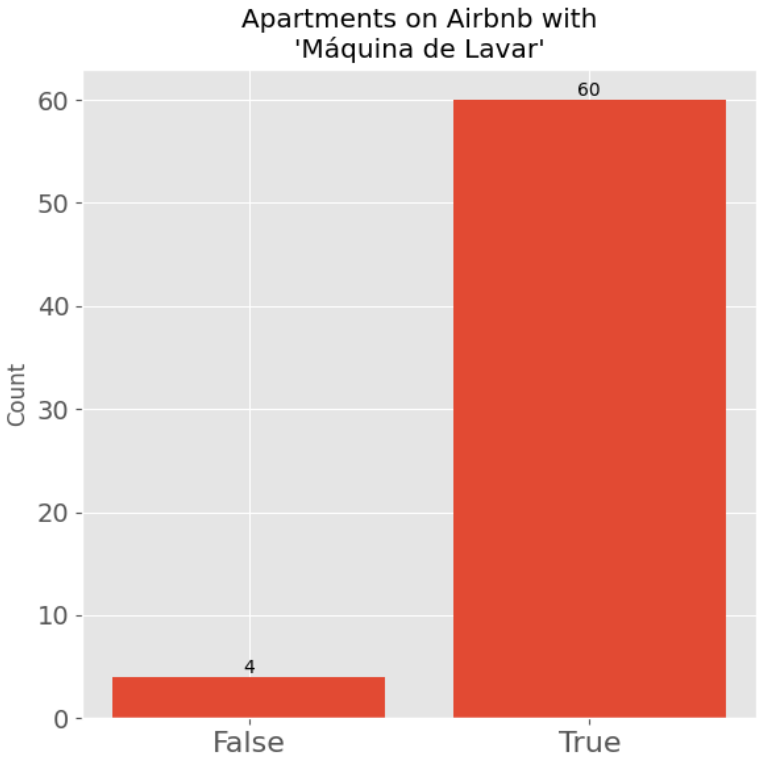
Having a refrigerator in the kitchen is almost a rule on top booked Airbnb apartments.



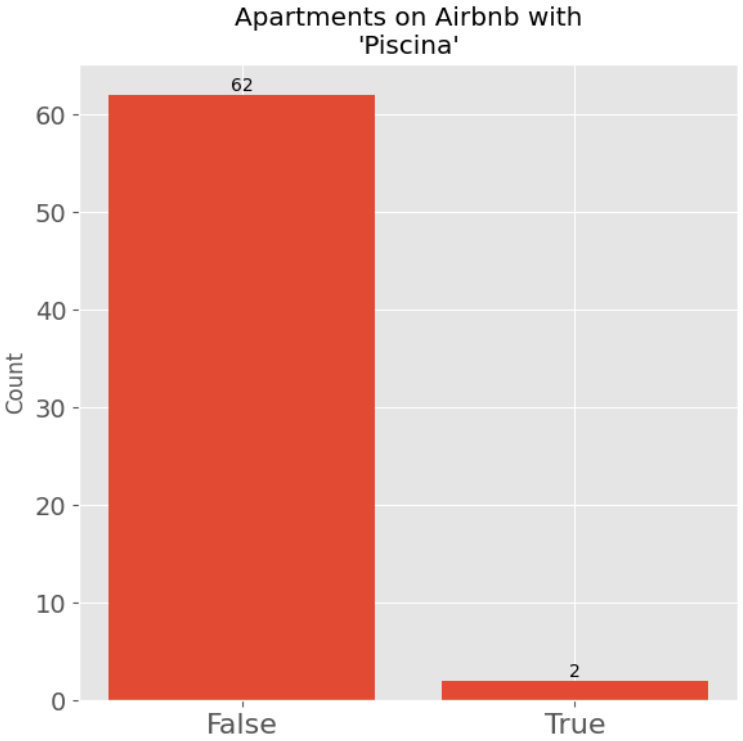
Having kitchen essentials is almost a rule on top booked Airbnb apartments.



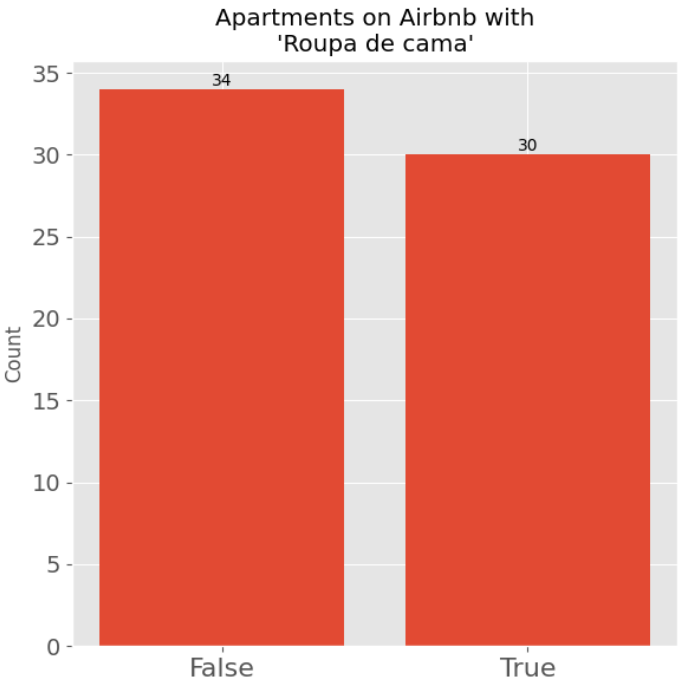
Beach essentials is not something that top booked Airbnb apartments have.



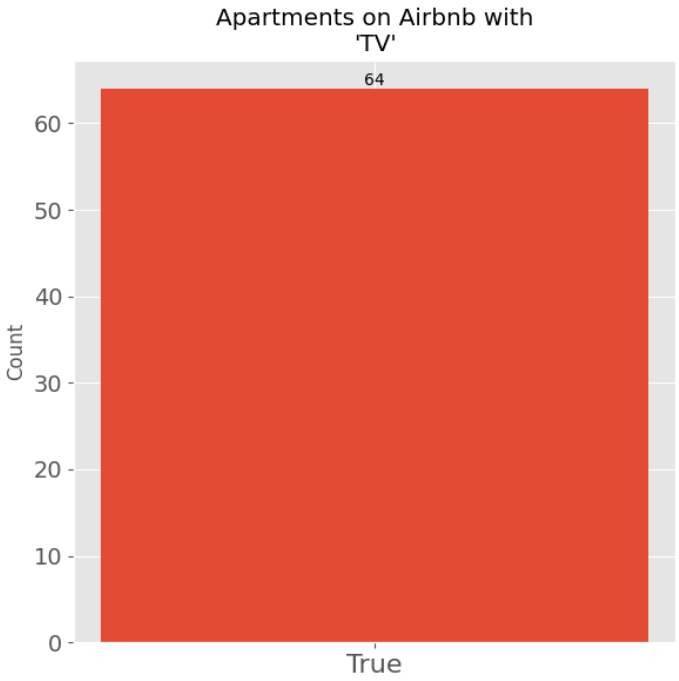
Having a washing machine is almost a rule on top booked Airbnb apartments.



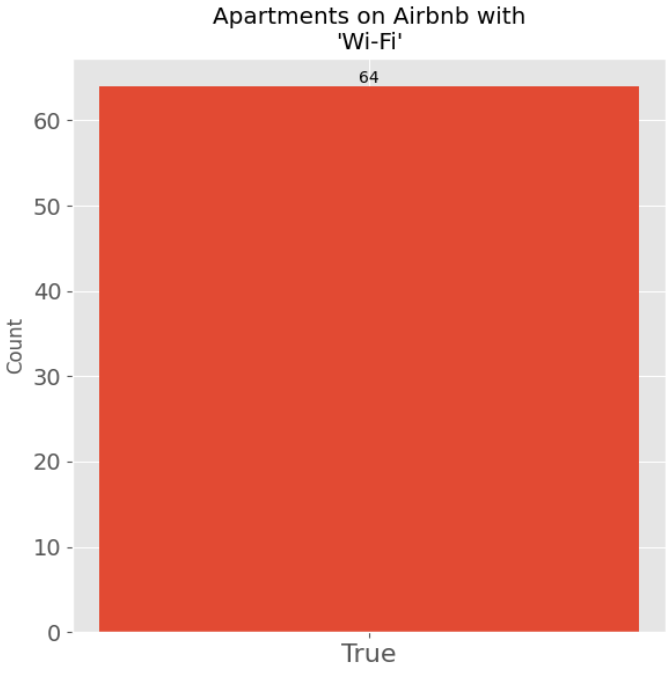
Having a pool is not something that top booked Airbnb apartments usually have. It makes sense, since all of these apartments are close to the sea.



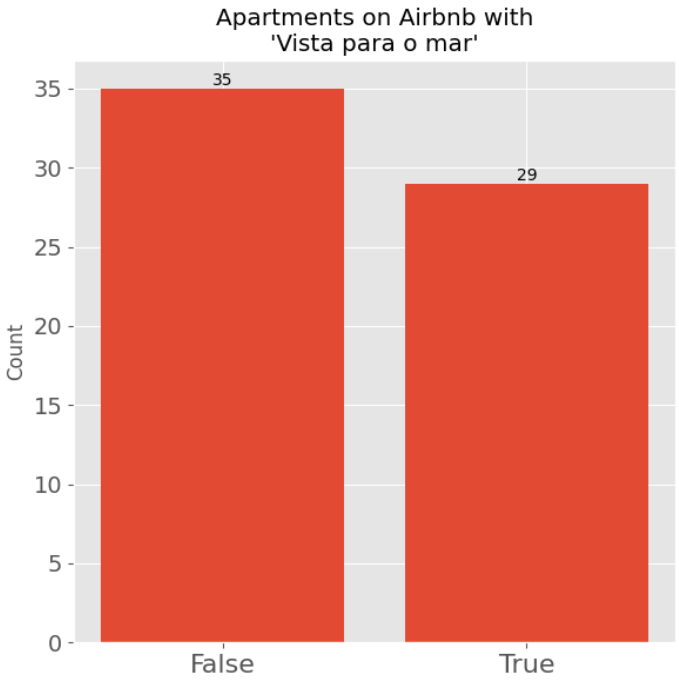
Having bedding is not a rule on top booked Airbnb apartments usually have.



As it was expected, all top booked Airbnb apartments have a tv.



As it was expected, all top booked Airbnb apartments have wifi.



Having a sea view is not a rule on top booked Airbnb apartments.

4.4. Question 4: We would like to build a building of 50 apartments in the city, where should we build it and how should the apartments be designed in order to be a great investment?

A “great” investment for a person thinking about buying a property would be one with a low vacancy and high revenues. These were already defined on questions 1, 2 and 3.

For Seazone it would also need to be considered how fast it could be sold. And there’s also another component: the price of the lot. Regions with low vacancy and high revenues naturally have more expensive lots, therefore it also needs to be analyzed in order to define where it should be built.

The following mean prices/m² calculated on VivaReal listings won’t differentiate if an entry has sea view or not, therefore they will only be useful to compare different regions, but they will not be used as a measure of a mean price of the region. For that, a manual inspection of VivaReal listings will be necessary for now.

Canto da Praia and Meia Praia have, each, 9 lots being sold, Centro has 10 and Morretes has the most lots being sold, 36.

While Canto da Praia and Meia Praia have the highest revenues, they also have the most expensive lots (R$10.109,00/m² and R$7.337,00/m², respectively). Morretes and Centro, on the other hand, also have a high revenue, but they have much cheaper lots (R$1.385,00/m² and R$3.184,00/m²).

Meia Praia has a lot of properties being sold (5.091), which could make selling apartments in the region hard. Morretes and Centro each have almost 5 times less properties than Meia Praia (1.716 and 1.116 respectively). Canto da Praia has only 60 properties being sold.

Canto da Praia and Centro have high priced apartments (R$16.897,00/m² and R$16.064,00/m²). Morretes and Meia Praia are currently being sold by a mean price/m² of R$9.814,00/m² and R$12.796,00/m², respectively.

Developing a 50 apartment building on Canto da Praia would be really expensive, and finding a good lot would also be really hard. Meia Praia is also an area really expensive to buy a lot, with many apartments being sold. They should not be considered.

Centro has an intermediate price for lots, high priced apartments and not many apartments being sold. Morretes is the cheapest to buy a lot, but it's not clear if they are near the sea or not. They are good candidates if a lot near the sea can be found.

But searching for lots in Morretes on VivaReal resulted only in really far from the sea properties, even in dirt roads! Therefore Morretes is not an option if a good lot cannot be found.

Centro doesn't have this problem, since the whole neighborhood is near the sea.

Therefore the recommended localization is the neighborhood Centro.

To summarize:

- **Profile (from question 1)**
  - **Apartment**
  - **Bedrooms: 2**
  - **Bathrooms: 1**
  - **Beds: 1 double bed, 1 single bed and a sofa bed**
- **Characteristics (from question 3):**
  - **No gym**
  - **Pets being allowed is unimportant**

    - **Has air conditioning**
    - **Has at least one parking spot (depending of price this could change)**
    - **Kitchen with:**
        - **Stove**
        - **Microwave**
        - **Refrigerator**
        - **Kitchen essentials**
    - **Having beach essentials is unimportant**
    - **Has a washing machine**
    - **No pool**
    - **Bedding is unimportant**
    - **Has a TV**
    - **Has wifi**
    - **Sea view is unimportant**
    - **Since it's a short term stay, furnitures should be present**
- **Location: Centro**

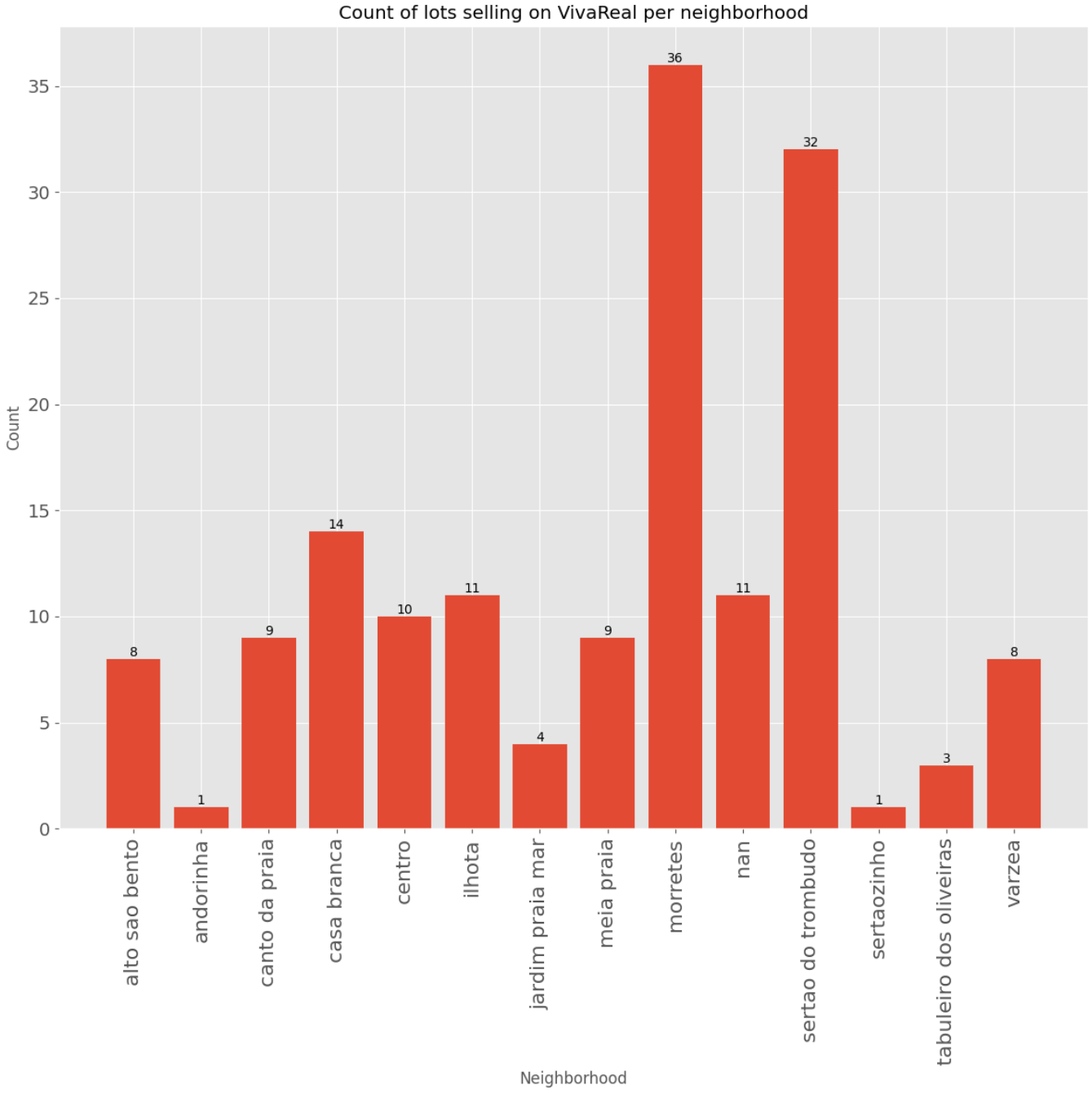
An idea of the apartment design is one with 35 m²:

Source: https://www.imovelweb.com.br/propriedades/saia-do-aluguel-02-dorms-no-iguatemi-2939780121.html

Below are the analyses made to come up to this conclusion.

4.4.1. Number of lots selling on VivaReal per neighborhood

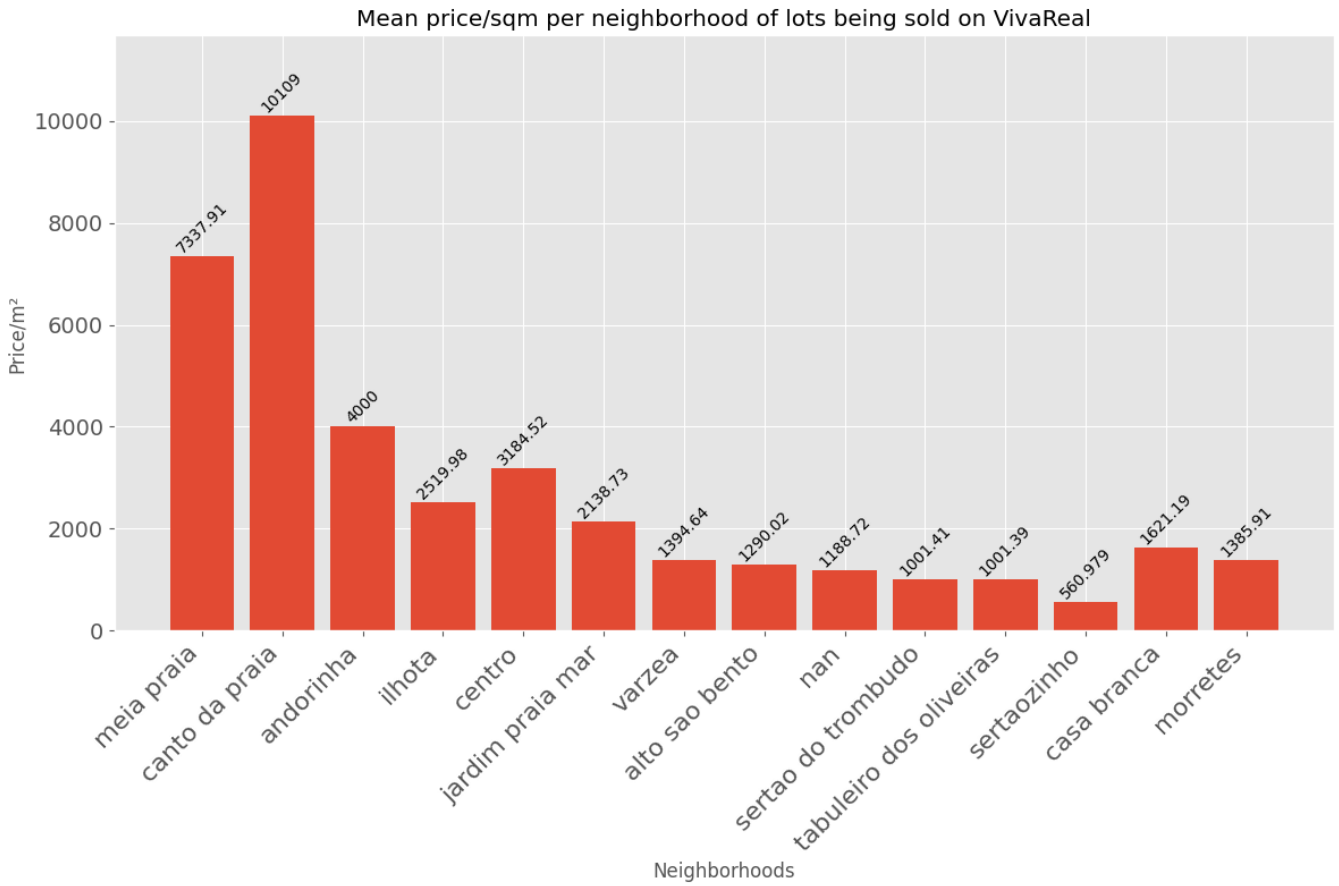
- dataframe = vivareal_lots_selling
- column = address_neighborhood



The neighborhoods of Morretes and Sertão do Trombudo have the most available lots selling on VivaReal. This can indicate a good opportunity! The next analyses confirms this theory.

4.4.2. Price/m² per neighborhood among VivaReal lots being sold

- dataframe = vivareal_lots_nbh
- column = address_neighborhood

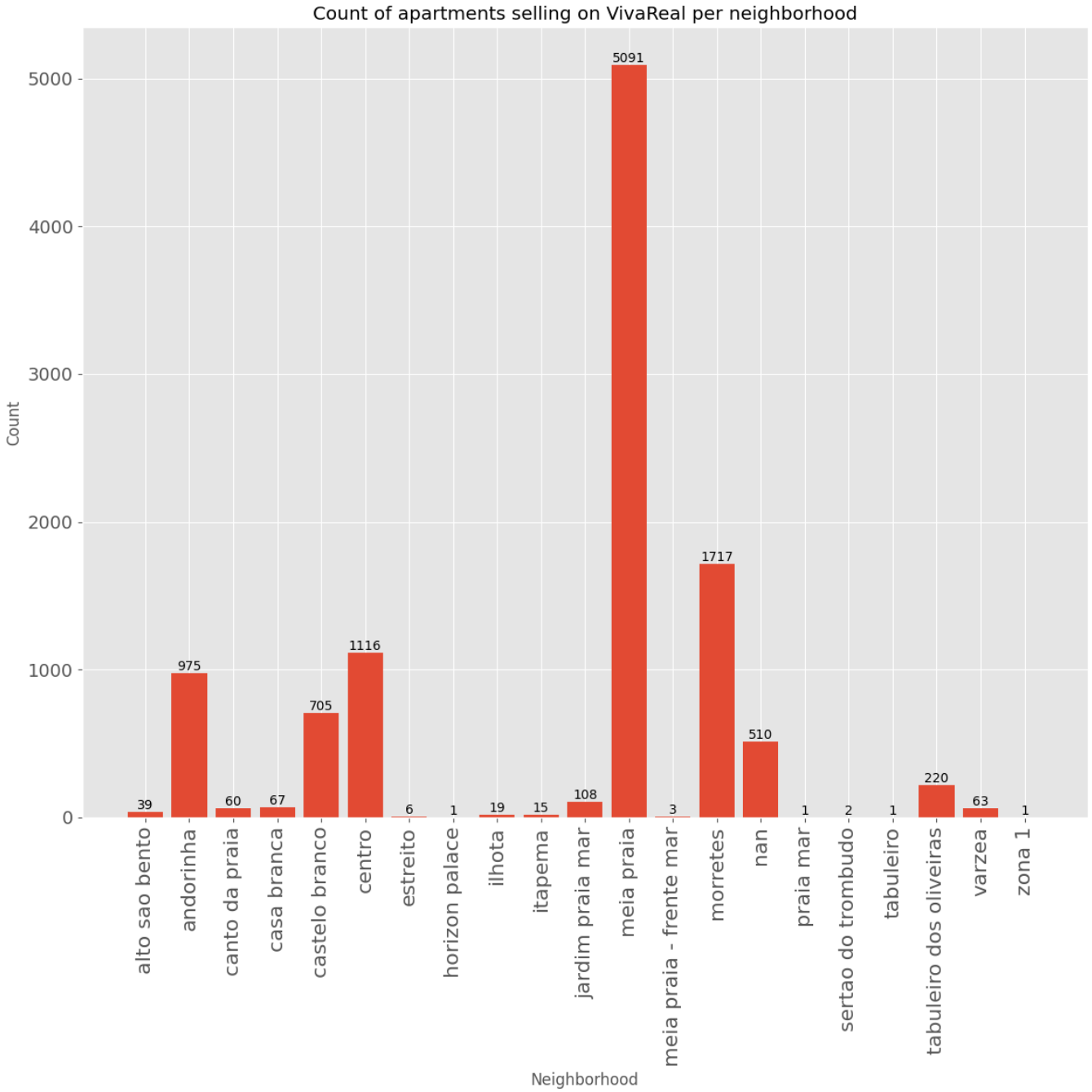


The neighborhoods of Morretes and Sertão do Trombudo have one of the lowest price/m² for lots among all neighborhoods on VivaReal.

Centro (downtown) is a region with an intermediate price/m² for lots, while Meia Praia and Canto da Praia have the highest.

4.4.3. Number of apartments selling on VivaReal per neighborhood

- dataframe = vivareal_apartments_selling
- column = address_neighborhood

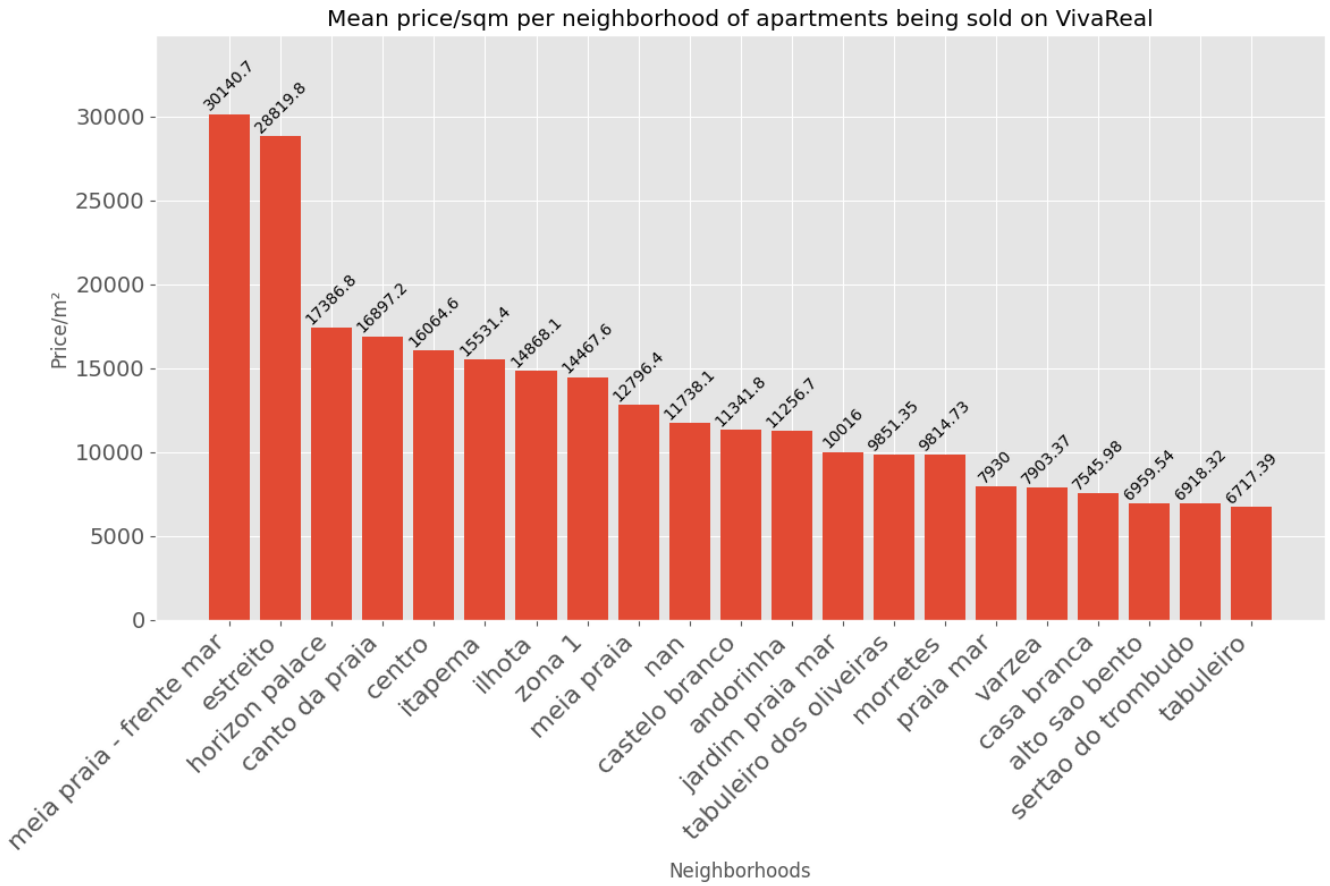


The region of Meia Praia has the most apartments being sold across VivaReal, 5.091 (but the next analysis indicates that its price is also one of the highest).

Morretes and Centro have a similar amount of properties being sold (1.717 and 1.116 respectively), while Canto da Praia has only 60 properties being sold!

4.4.4. Price/m² among VivaReal apartments being sold

- dataframe = vivareal_apartments_nbh
- column = address_neighborhood



“Meia Praia - frente mar” is the most expensive neighborhood (R$30.140,00/m²) for apartments. Canto da Praia and Centro also have high prices (R$16.897,00/m² and R$16.064,00/m²).

Morretes and Meia Praia (not to be confused with Meia Praia - frente mar) are currently being sold by a mean price/m² of R$9.814,00/m² and R$12.796,00/m², respectively

4.5. Question 5: How much will be the return on investment of this building in the years 2024, 2025 and 2026?

It's important to notice that if a lot were to be bought and a building developed on it, the construction would probably be completed around 2026 (there are design phases, approval phases at the city hall and only then construction).

Hence the return on investment would not make sense to calculate in the years 2024, 2025 and 2026. Therefore I am considering what the ROI will be when the apartments are completed (around April 2026).

Net Present Value (NPV) is a great indicator of a good investment, and it was also calculated.

Three scenarios were created: selling at a normal price, 110% of the price (good) and 90% of the price (bad).

ROI of this building, considering that it would be done by April 2026, will be:

- **21,79% in the bad scenario**
- **33,48% in the normal scenario**
- **45,16% in the good scenario**

NPV of this building will be:

- **- R$107.899,91 (a negative NPV indicates that it's better to invest in something else)**
- **R$585.295,05**
- **R$1.278,490,01**

Using the support Excel spreadsheet[4], the following calculations were made (on next page):

[4] Sorry for changing to Portuguese, it was created a long time ago in this language

INFOS DO TERRENO	
Endereço	SC - Itapema
Bairro	Centro
Zoneamento	ZAP-1
Qualificação Ambiental	n/a
Área terreno	980,00 m²
CA mínimo	0,15
CA básico	1
CA máximo	2
Valor venal	n/a
Recuo lateral (após 10m de altura)	n/a

INFOS DO PROJETO	VALOR
Área construída (AC)	100,00%
Área CA básico	1960,00 m²
Área CA máximo	980,00 m²
Custo outorga onerosa	1960,00 m²
Taxa de ocupação (TO)	
Área ocupável máxima	80%
Altura máxima	784,00 m²

INFOS DO EMPREENDIMENTO	VALOR
Andares	3
Pé-direito interno de cada andar	2,8 m
Altura do prédio	9,06 m
Área vendável	1764,00 m²
Aptos por andar	16,66666667
Número total de aptos	50
Área de cada apto (média)	35,00 m²
Tipo projeto	R-16 - Padrão de acabamento: Normal
CUB - Custo Unitário Básico	R$ 2.281,45 /m²

RESUMO CUSTOS			
TIPO	PERCENTUAL		VALOR
Custo terreno		R$	3.120.320,00
Custo construção		R$	4.667.642,00
Custo projetos		R$	294.000,00
Custo burocracia		R$	55.584,71
Impostos	6,73% do VGV	R$	883.312,50
Comissão corretores	5% do VGV	R$	656.250,00
Marketing	1,0% do VGV	R$	131.250,00
CUSTO TOTAL		**R$**	**9.808.359,21**

CURVA DE VENDAS				
PERÍODO	% DE APTOS VENDIDOS	N° DE APTOS VENDIDOS	% PAGA NA CONSTRUÇÃO	% PAGA NAS CHAVES
Lançamento	50%	25 un.	100%	0%
Durante as obras	30%	15 un.	100%	0%
Entrega das chaves	5%	2,5 un.	0%	100%
até 1 ano após entrega das chaves	10%	5 un.	n/a	n/a
entre 1 e 2 anos após entrega das chaves	5%	2,5 un.	n/a	n/a

DATAS E TEMPOS		
PERÍODO	MESES	DATA DO FIM DO PERÍODO
1º aporte	-	2/1/2023
Tempo pré-aprovação	3	5/1/2023
Projeto/Aprovações	6	11/1/2023
Lançamento	5	4/1/2024
Obras	24	4/1/2026

RESUMO VENDAS	VALOR - CENÁRIO NEGATIVO	VALOR - CENÁRIO NORMAL	VALOR - CENÁRIO POSITIVO
	90%	100%	110%
Preço de venda (/m²)	R$ 6.750,00	R$ 7.500,00	R$ 8.250,00
Preço de venda do apto	R$ 236.250,00	R$ 262.500,00	R$ 288.750,00
VGV	R$ 11.812.500,00	R$ 13.125.000,00	R$ 14.437.500,00

INDICADORES	VALOR - CENÁRIO NEGATIVO	VALOR - CENÁRIO NORMAL	VALOR - CENÁRIO POSITIVO
Lucro	R$ 2.137.651,73	R$ 3.283.562,67	R$ 4.429.473,60
Tir (a.m)	1,670% a.m.	2,411% a.m.	3,071% a.m.
Tir (a.a)	21,994% a.a.	33,087% a.a.	43,768% a.a.
ROI (margem sobre investimento)	21,79%	33,48%	45,16%
VPL	**-R$ 107.899,91**	**R$ 585.295,05**	**R$ 1.278.490,01**

Taxa básica de juros	13,75% a.a.
+	
Taxa extra do investidor	10,00% a.a.
=	
TMA	1,792% a.m.
TMA	23,750% a.a.

5. Future work and feedback

There is still a lot of things that can improve in different parts of the project to create better analyses, for example (non-exhaustive list):

1. Double check poorly formatted strings (and check the scraper)
2. Use Apache Spark to increase analyses speed
3. Many listings (especially on VivaReal), though having different ids, are actually the same property, only announced by a different real estate agency. It may be possible to identify these based on the pictures and then drop the duplicates.
4. A lot of information can be extracted from a listing textual description that the owner provides, for example size of the property, age, distance to the sea and many more.
5. Age of the property may also be inferred using Machine Learning (given that a big labeled database of listing pictures is available)
6. Group VivaReal listings with and without sea view, to create better mean price/m² estimates
7. Repeat analysis for each neighborhood to discover what is the profile and characteristics of apartments on each neighborhood (instead of doing for the top booked apartments across the whole city)
8. Check if there's any way to get latitude and longitude from VivaReal listings
9. Use Uber's H3 Hierarchical Spatial Index instead of geohash and check if it makes any difference
10. Utilize Airbnb Hosts data (it wasn't utilized in this project)

I am really amazed by how much fun it was to complete this challenge.Working with maps is really beautiful. It's a really interesting problem: even though we are talking about one small city, the opportunities that all of the scraped data creates are huge. There's a lot to cover, explore and many different approaches to "solve" the challenge.

I also learned different things when tackling the problems and really appreciated the effort put to create the datasets and formulating the questions.

6. Questions answers

6.1. Question 1: What is the best property profile to invest in the city?

Property profile 1:

- Apartment
- Bedrooms: 2
- Bathrooms: 1
- Beds: 2 double beds
- Location: see question 2
- Others characteristics: see question 3

6.2. Question 2: Which is the best location in the city in terms of revenue?

The best locations (neighborhood) in terms of revenue are Canto da Praia and Meia Praia, followed by Centro and Morretes (this one considering only properties near the sea).

6.3. Question 3: What are the characteristics and reasons for the best revenues in the city?

Characteristics that top booked airbnb apartments (with the highest revenues) have are:

- Profile
    - Apartment
    - Bedrooms: 2 or 3
    - Bathrooms: 1 or 2
    - Beds: 2 to 4 double beds (to accommodate up to 6 guests)
- Location: Canto da Praia, Meia Praia, Morretes and Centro
- Characteristics:
    - No gym
    - Pets being allowed is unimportant
    - Has air conditioning
    - Has at least one parking spot
    - Kitchen with:
        - Stove
        - Microwave
        - Refrigerator
        - Kitchen essentials
    - Having beach essentials is unimportant
    - Has a washing machine

    - No pool
    - Bedding is unimportant
    - Has a TV
    - Has wifi
    - Sea view is unimportant
    - Since it’s a short term stay, furnitures are always present

6.4. Question 4: We would like to build a building of 50 apartments in the city, where should we build it and how should the apartments be designed in order to be a great investment?

To summarize:

- Profile (from question 1)
    - Apartment
    - Bedrooms: 2
    - Bathrooms: 1
    - Beds: 1 double bed, 1 single bed and a sofa bed
- Characteristics (from question 3):
    - No gym
    - Pets being allowed is unimportant
    - Has air conditioning
    - Has at least one parking spot (depending of price this could change)
    - Kitchen with:
        - Stove
        - Microwave
        - Refrigerator
        - Kitchen essentials
    - Having beach essentials is unimportant
    - Has a washing machine
    - No pool
    - Bedding is unimportant
    - Has a TV
    - Has wifi
    - Sea view is unimportant
    - Since it’s a short term stay, furnitures should be present
- Location: Centro

6.5. Question 5: How much will be the return on investment of this building in the years 2024, 2025 and 2026?

ROI of this building, considering that it would be done by April 2026, will be:

- 21,79% in the bad scenario
- 33,48% in the normal scenario
- 45,16% in the good scenario

NPV of this building will be:

- - R$107.899,91 (a negative NPV indicates that it’s better to invest in something else)
- R$585.295,05
- R$1.278,490,01