Seazone Challenge - Report

Introduction

This document has the goal to report the execution of seazone challenge as required. I will answer here the questions raised also describing problems tackled, solutions and simplifications chosen.

This is a personal report of results and reflections on tackling the problem, then some informal tone should be expected.

Technology stack

In order to tackle the problem I chose to use Python 3 with pandas (or dask) to structure the solution. That choice sounds natural or obvious to me because it's mainly a data science and data engineering challenge. Although I know the technology and did sporadic tasks with that, I'm not on my toes with this technology. Still the choice is the best because I'd have to deal with large datasets, some unstructured data and add some rich combinations producing some code. I could analyze with Google Big Query or even Excel, but I'd not produce visible code very easily on those platforms.

Datasets used

I focused on answering the questions made. Because of that I decided not to use the hosts dataset. Also, in order to haste and produce solutions quickly I decided not to use the VivaReal dataset. Of course for a detailed and complete analysis it should be used. In order to use that it'll be necessary to reconcile categories like amenities and others to make a complete analysis. Also VivaReal has more important data like property sizes that should be considered in the analysis. I don't know why the AirBnb datasets don't have this data so it wasn't considered in the analysis. Anyway, for a complete analysis I believe it's very important information.

Data processing and outputs

Data preparation and pre-analysis

The prices database is pretty huge but it's the source for revenues, so I did a data prep before by consolidating this database by listing and year and only by listing. These will generate the files Price_grouped.csv and Price_grouped_Year.csv, both will be placed in the datasets folder. The dataset contains data for several acquisition dates. So, I decided to pick only one price and disregard all other dates. The proper thing to do I believe would be to pick every listing on the same acquisition date or pick the newest of each listing. I had difficulties doing that because acquisition dates are datetimes and I'm not secure of how this date is produced. Also I couldn't filter the data by date (without time) of each listing. So, for simplicity and for demonstration purposes I've just discarded duplicates, however for a clear and concise analysis it shouldn't be made.

Those datasets contain the prices accumulated by each listing (either by year or only the listing). For the accumulation I considered only the listings prices when they were not available because I interpreted that most of the time it wasn't available because it was occupied and generating revenue. I know that unavailability could have other reasons like the host is using the property. As I could not find any data that demonstrates that this was the case, then I decided to consider it as occupied by guests. Possibly that's some kind of market default for this situation (like 70% of cases of unavailability it's because it's occupied by guests) but I couldn't find one. At least if it's wrong I expect that everything is wrong on the same dimension which affects the analysis less (except for the ROI).

Output folder

All outputs generated for analysis purposes are placed in the output folder under the project folder.

Clusterization

With revenues by listing I used the geolocation dataset to clusterize the listings. Initially I tried to get the neighborhoods of each listing by using nominating, but I believe that was too much data and I was blocked by their API. So, then, I did a generic clusterization and analyzed localizations visually. It generates a html file with clusters numbers centered on the map, after execution the file resides on clusters.html.

With clusters generated I analyze the revenue and quantity of listings of each cluster. So it generates two graphs with this analysis estimated_revenues_by_cluster_year.png and number_listings_by_cluster_year.png.

Listing traits and profile

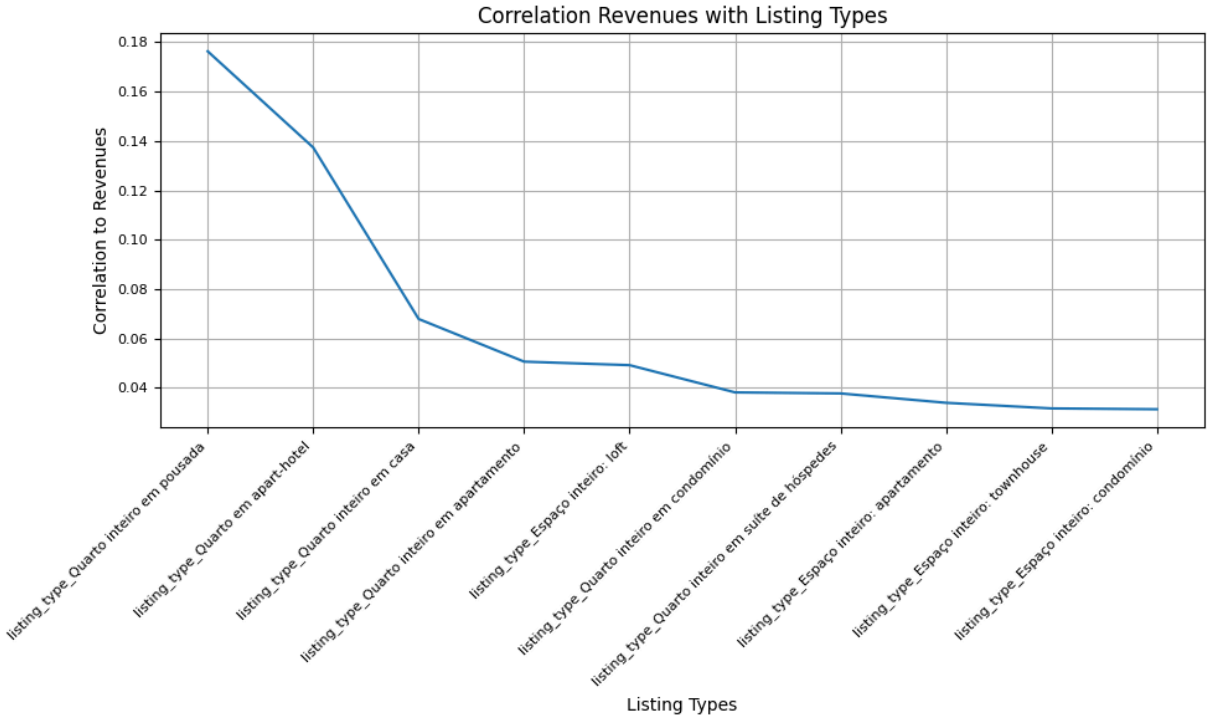
For analyzing the possible traits (or characteristics) of each listing and evaluating how each can affect revenue flow I correlated the estimated revenue with each trait considered. First it evaluates against every numeric variable in the Details dataset generating traits.png graph. Then it extracts safety features, house rules and amenities fields, categorizes them and generates for each one a graph with correlations. This extraction was tricky because despite the fields being in the same structure it was a non-canonical structure. So I had to use a regular expression to parse that into a list and, then, into categories. Finally it categorizes the listing_type field and correlates that also with estimated revenue generating the graph in listing_types.png

Answering question and analysis

Here, I'll list all questions raised in the challenge and answer each one based on the data generated.

What is the best property profile to invest in the city?

I considered profile as the listing types because it's an overall scout of properties in the city. Therefore I used the listing types analysis in order to answer this question:

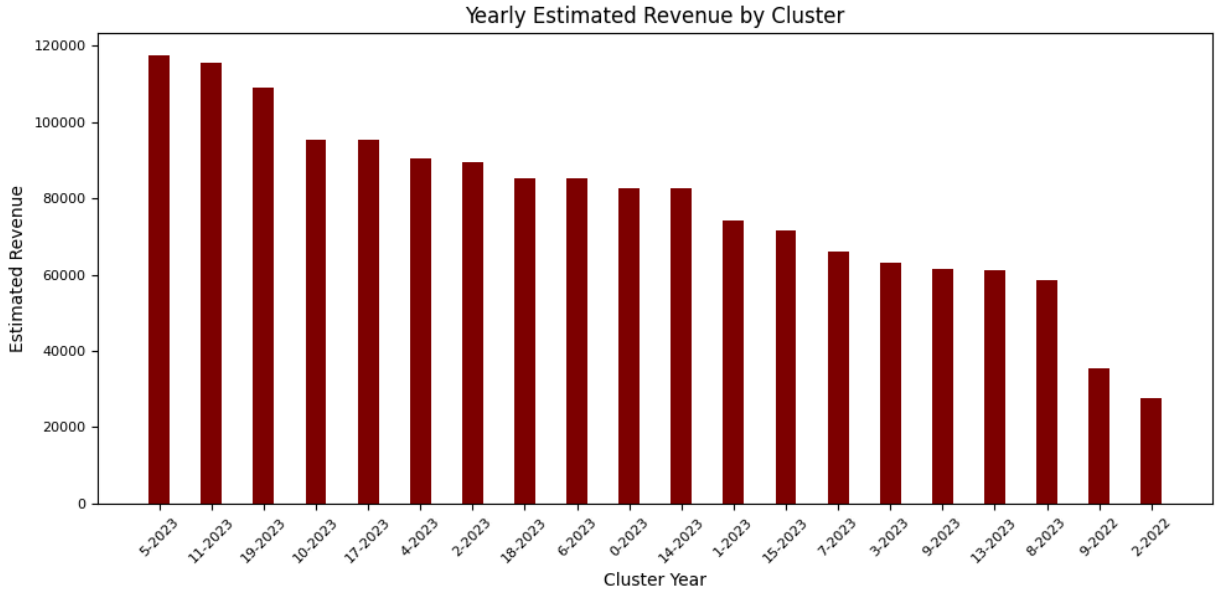


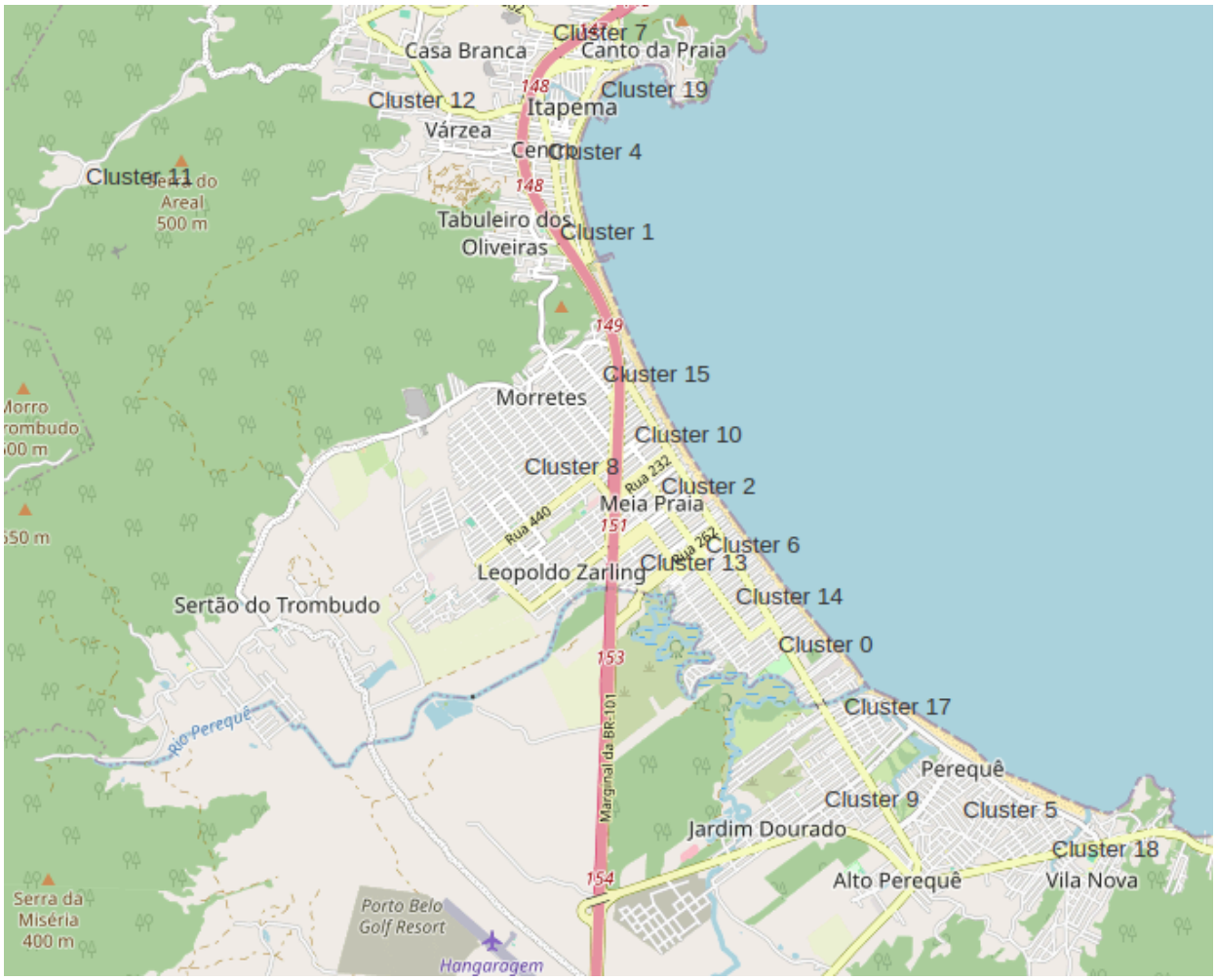
By looking into the graph the guesthouses are the best property profiles to invest in. With a bit more detail, by analyzing the amenities graph, if the guesthouse has a sea view it'll be much more valued by potential guests.

Which is the best location in the city in terms of revenue?

By analyzing the data provided, I consider the region of **Perequê.**

That's because considering an analysis of 20 clusters in the city the estimated revenue accumulated per cluster is well distributed in the city:





Despite having a high revenue, the cluster 11 has very few listings, so even though it generates a high revenue it must be considered it lacks options.

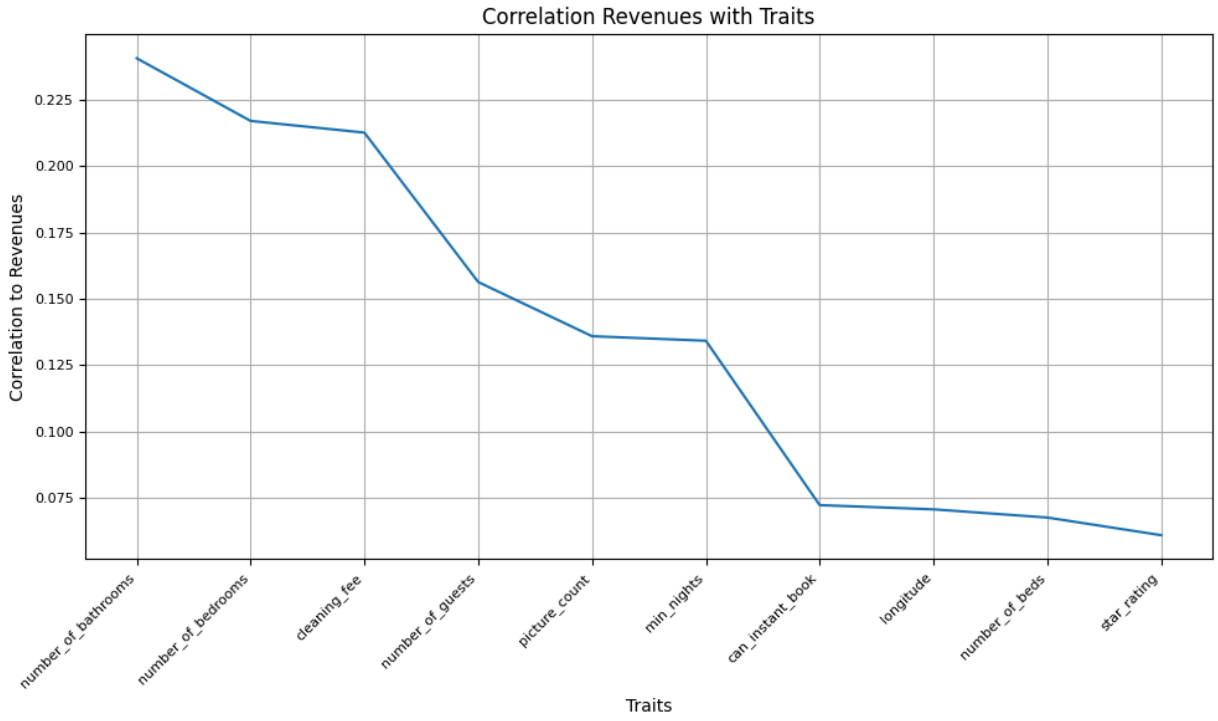
Considering that clusters 5, 10 and 19 are on different locations in the city and also that cluster 5 has the highest revenue and is close to cluster 17 which has also a high revenue, then Perequê should be the best location

What are the characteristics and reasons for the best revenues in the city?

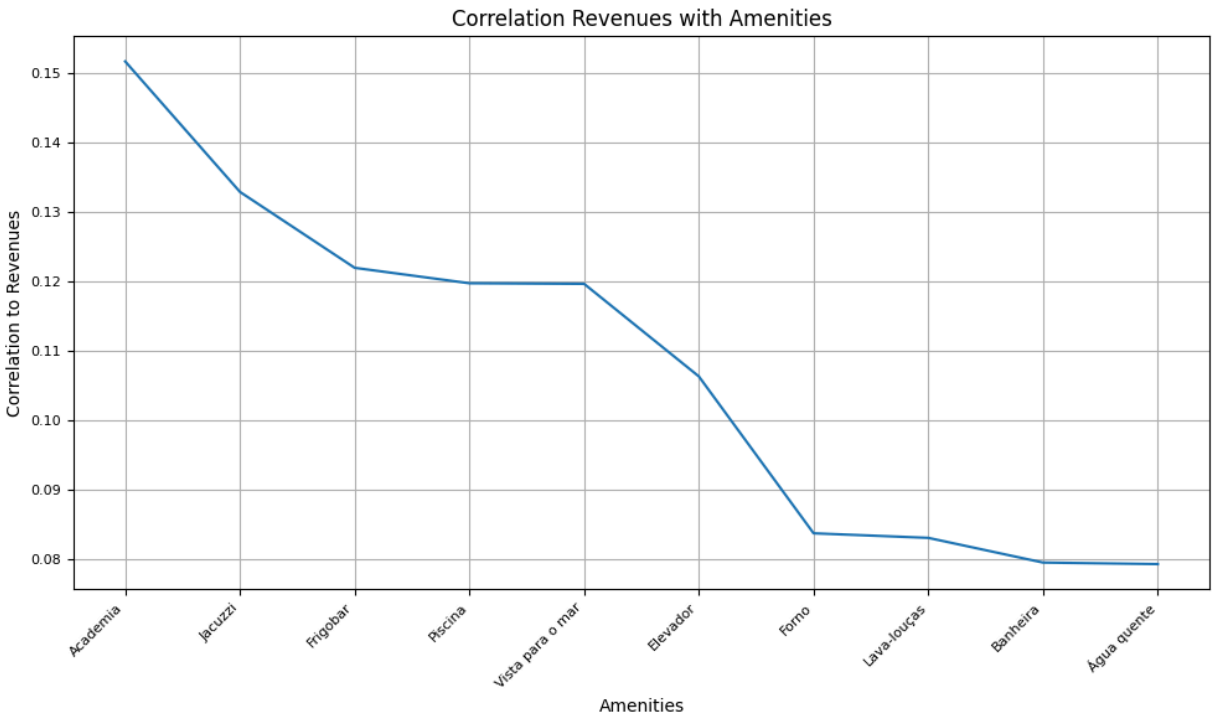
To achieve higher revenues the best is to invest on larger properties with features such as academies, jacuzzis and pools. Those features are more appropriate for houses or apartments but seahouses also generate very good revenues. Either of them the seaview is an important characteristic which seems to be valued by guests. Also, safety features such as locks in the room and facilities such as self check-in seem valuable. Some strict rules such as unallowing parties and events and prohibiting smoking also look to increase revenues. Finally, some

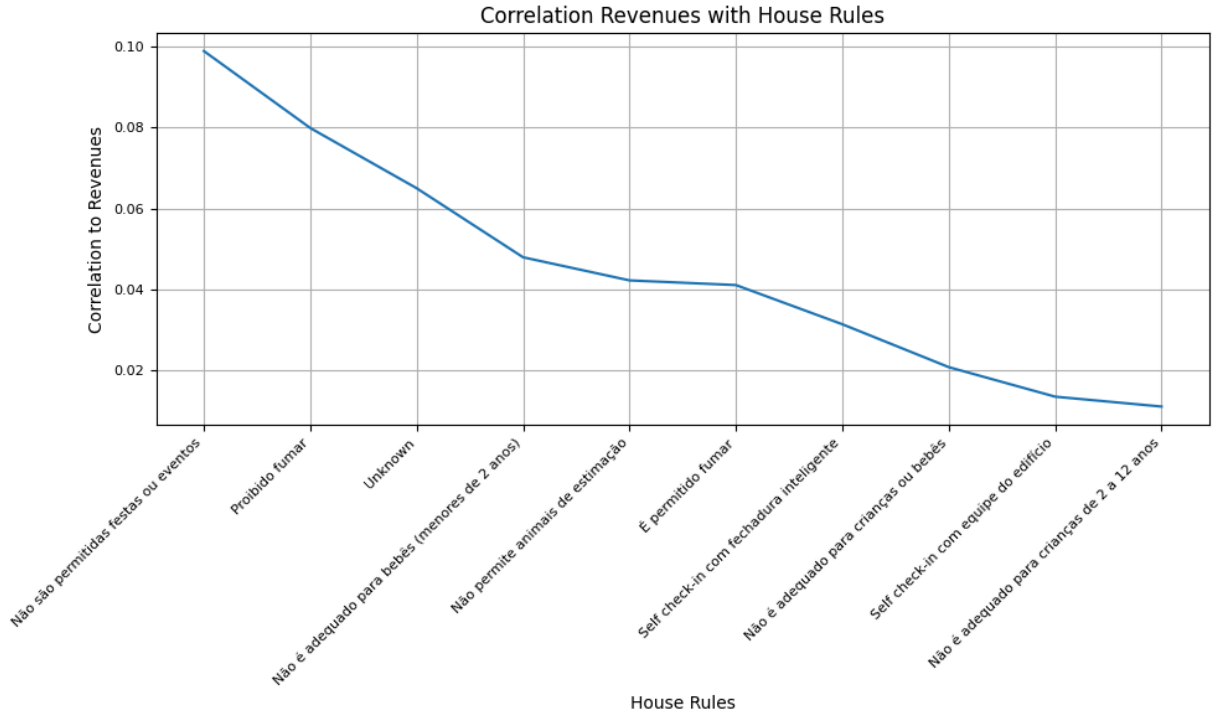
features to the listing such as adding more pictures and allowing instant booking must attract more guests.

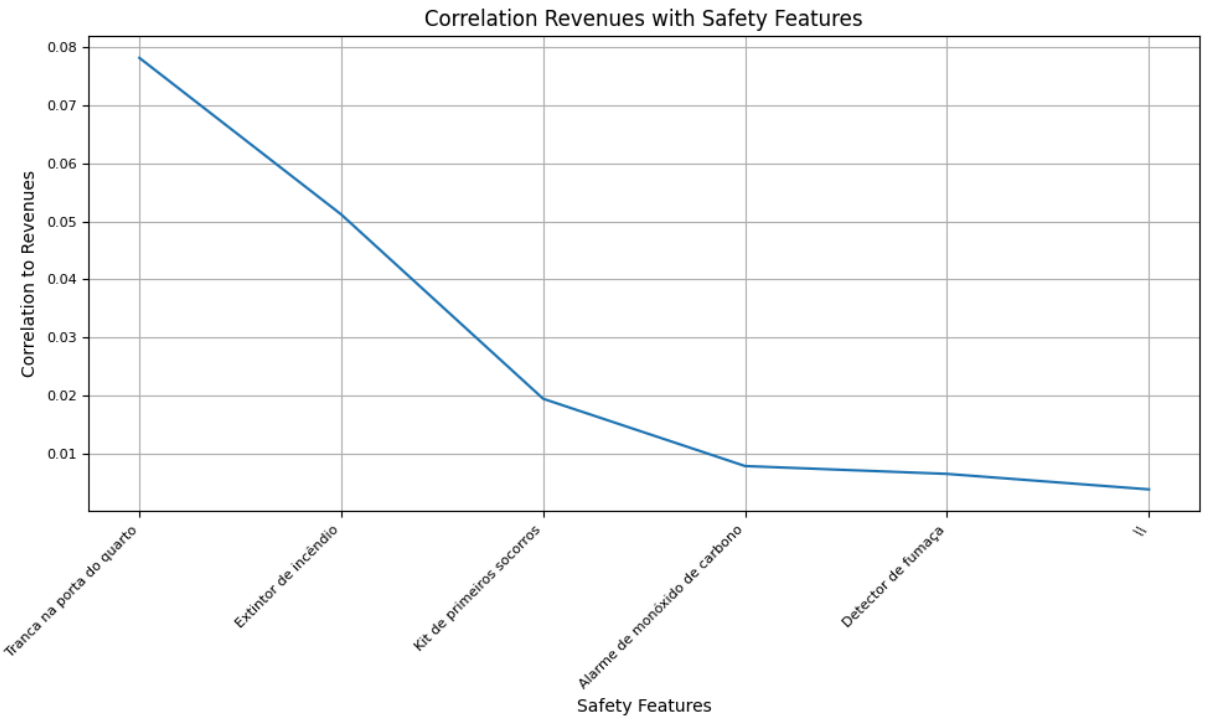
This analysis considered some correlations as traits. This one specifically in order to understand that larger properties generate higher revenues because the number of bathrooms, bedrooms and guests increase revenues, I believe that those traits apply to larger properties. This analysis also brought insights regarding listing features as count of pictures and instant booking.



Also, the analysis of amenities, house rules and safety features improved the overall characteristics of listings:


Correlation Revenues with Amenities
Correlation to Revenues
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
Academia
Jacuzzi
Frigobar
Piscina
Vista para o mar
Elevador
Forno
Lava-louças
Banheira
Água quente
Amenities


Correlation Revenues with House Rules
Correlation to Revenues
0.10
0.08
0.06
0.04
0.02
Não são permitidas festas ou eventos
Proibido fumar
Unknown
Não é adequado para bebês (menores de 2 anos)
Não permite animais de estimação
É permitido fumar
Self check-in com fechadura inteligente
Não é adequado para crianças ou bebês
Self check-in com equipe do edifício
Não é adequado para crianças de 2 a 12 anos
House Rules



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?

The ideal would be an apart hotel in the region of Perequê with traits as academy, pools and near to the shore, which means with seaview. It should have locks in the rooms and pay attention to safety features such as first-aid kits, smoke detectors and fire extinguishers regardless of that possibly those are mandatory.

How much will be the return on investment of this building in the years 2024, 2025 and 2026?

Ok, in order to evaluate the return on investment it must consider the costs. I have any knowledge either on construction or on the region. Then I asked ChatGPT to calculate the costs for me. Here's the analysis it generated which I considered pretty reasonable:

> *To calculate the construction cost considering average finishing, we can make some estimates based on the average construction costs in the region. I'll provide an approximate estimate based on general data, but it's important to remember that actual costs may vary depending on various specific project and local market factors. Here's a general estimate:*

Labor Cost: In Santa Catarina, the average cost of labor in the construction industry can range from R$ 250.00 to R$ 400.00 per square meter. Let's use an intermediate value of R$ 300.00 per square meter.

Materials Cost: The cost of construction materials can vary significantly depending on the type and quality of materials chosen. Let's use an average value of R$ 700.00 per square meter.

Earthmoving and Foundation Cost: For earthmoving and foundation, estimate about 10% to 15% of the total construction cost. Let's use 12% of the total construction cost.

Based on these estimates, we can calculate the total construction cost:

Calculation of Total Built Area: As mentioned earlier, the total built area is 2500m².

Labor Cost: 2500m² x R$ 300.00/m² = R$ 750,000.00.

Materials Cost: 2500m² x R$ 700.00/m² = R$ 1,750,000.00.

Earthmoving and Foundation Cost: 12% x (R$ 750,000.00 + R$ 1,750,000.00) = R$ 270,000.00.

Total Construction Cost: R$ 750,000.00 (labor) + R$ 1,750,000.00 (materials) + R$ 270,000.00 (earthmoving and foundation) = R$ 2,770,000.00.

That being so, I consider the costs to be R$2.7M. Then it's necessary to collect the maintenance costs yearly. I guessed a 5% yearly maintenance cost, which represents R$138.5K / year. So, for the revenues I looked at that those apartments in AirBnb and those round around R$400 / day as the price of booking and as a conservative prediction I consider that each apartment will be occupied a bit more than 50% of days in a year, rounding-up it to 200 days a year. That goes with R$80K / year / apartment. By looking at the cluster analysis, the price median in the cluster 5 is of R$79.4K which aligns to this analysis.

In that analysis the revenue yearly is R$80k * 50 = R$4M gross revenue, discounting the costs of R$138.5K it results in a **net revenue of R$3.86M** a year, Considering this revenue the ROI is in 0.69 years which is **between 8 and 9 months**.

I consider this analysis very optimistic for the region. Notwithstanding, if the revenues are cut by half, the ROI is in less than 2 years which is a pretty good investment.

Post-analysis and feedback

This analysis considered the data provided and I did the best I could to produce a good analysis. Some more resources could be used in order to improve the analysis. Machine learning models could group more the characteristics and produce insights with all data. With

more information such as property sizes and more detailed booking information would generate better revenue estimating. At least I hope that any mistakes go all in the same direction, and that this analysis could be roughly accurate.

The challenge is interesting in order to evaluate the skills it intends. There are really necessary data wrangling and analysis skills here. Conversely, I considered the challenge pretty extensive. That is some pretty hard datasets and so many questions to answer. Because of that I had to simplify a few things and discard some datasets. Less questions and a few less, even difficult, datasets would be sufficient to evaluate the skills.