Challenge 1 - Data

Introduction

Some treatments were carried out on the available database, with the respective data sets:

- Removal of prices that represent outliers under analysis
- Removal of duplication in details file
- Generation of reduced price file
- CSV file generation with the possibility of evaluating the number of rentals, average price and revenue.

About the price file

It was carried out work on the file Price_AV_Itapema-001.csv, where the available dates were disregarded and only dates already rented were considered. Treatment was carried out regarding the overlapping of dates when the acquisition date is changed and the dates at that time onwards they are all recreated. From the treatment carried out, the file reduced from more than 40 million lines to around 400 thousand, this treatment and new file (Price_AV_Itapema_filtered) are generated by the Python algorithm.

Identification of outliers in average prices. Prices greater than or equal to 5000 were disregarded, as they caused significant distortions in average prices, therefore the figure below shows the best adjusted distribution, after disregarding the respective prices. All other analyzes were considered with this same view.

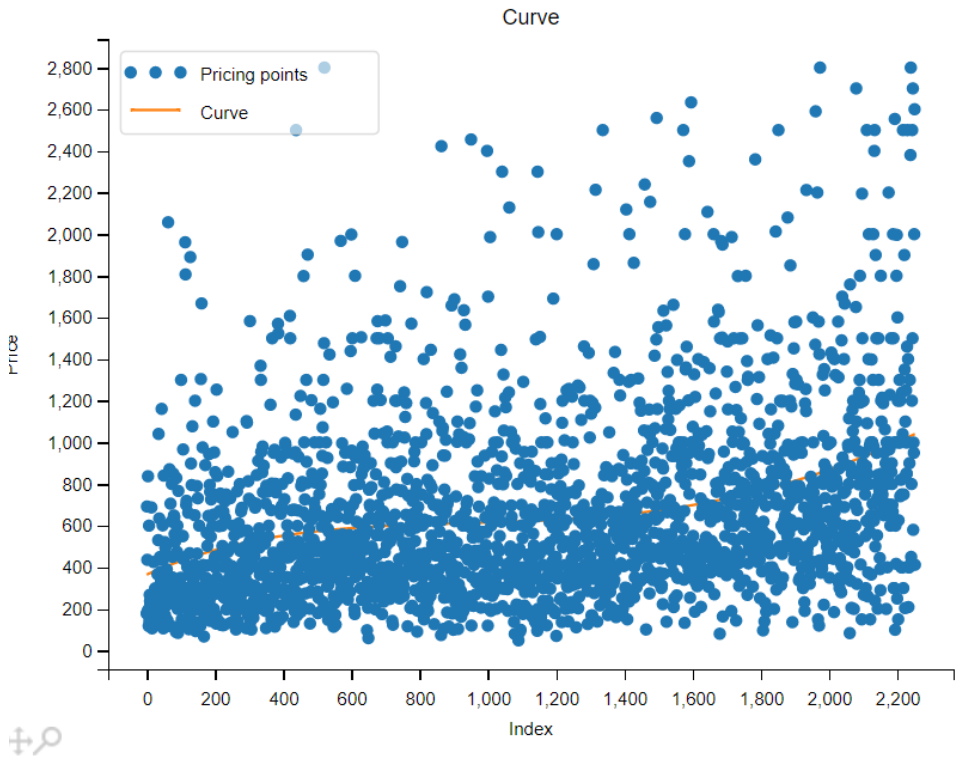

Figure 1 – Distribution of average prices

1. **What is the best property profile to invest in the city?**

To answer this question, a file called Property_profile.csv was generated, where details of rentals with average price and number of rentals are grouped.

The best profile to invest in is full space apartments, with around 75% of apartments available for rent compared to a group of approximately 2274 properties.

It is worth mentioning that the top 2 on the list of most rented are 2 identifiers referring to houses and these in turn are in smaller quantity, around 354 houses in the region alone for rent.

ad_id	listing_type	average_price	number_rented	revenue
39132422	Espaço inteiro: casa	178.14	549	97800.00
47018336	Espaço inteiro: casa	691.95	548	379191.00
12490354	Espaço inteiro: apartamento	436.05	544	237213.00
46088002	Espaço inteiro: apartamento	188.54	539	101623.00
52666674	Espaço inteiro: casa	837.20	533	446229.00

Figure 2 – Greater number of rentals

2. **Which is the best location in the city in terms of revenue?**

Close to Orla Meia Praia, according to the image generated by the 30 locations that represent revenues, approximately, above 500 thousand during the evaluated data period. Map is generated by code with the name mapa_itapema.html

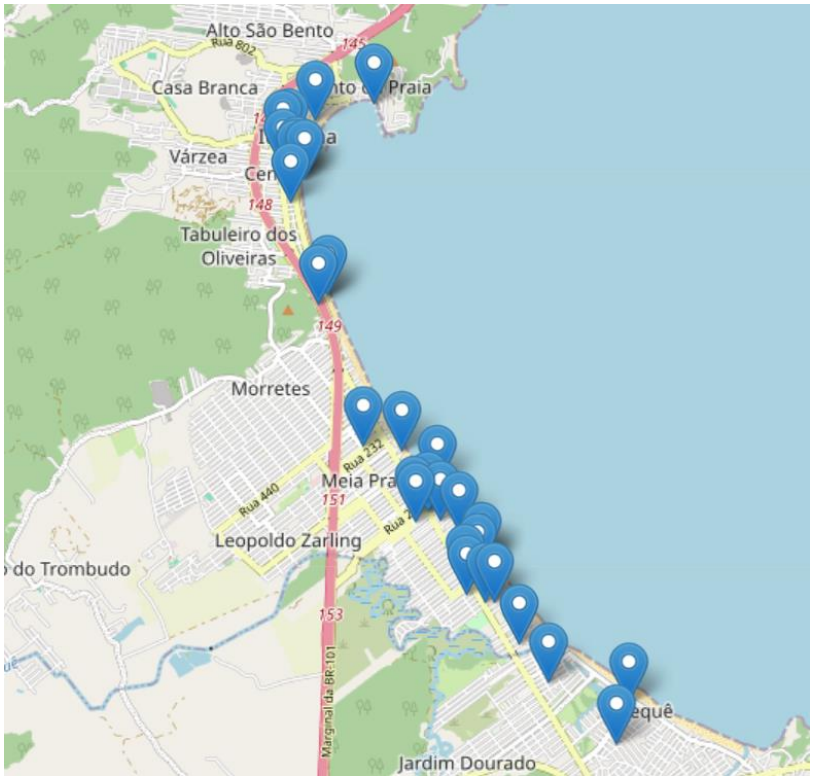

Figure 3 – Map with distribution of the 30 most rented places

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

Close the Orla da Meia Praia region there are several businesses in different sectors, such as: supermarket, restaurants (examples: Macdonalds, Burger King), coffee shop (Kopenhagem), pharmacies, banks, shopping malls, etc.; This ensures that guests are provided with most of their day-to-day needs. For example, the locations highlighted by the rectangle in the image below.

Figure 4 – Zoom in - Local commerce near Orla Meia Praia

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?**

Close to the beachfront of Meia Praia, as previously shown and the apartments must have 3 to 4 bedrooms, 3 to 4 bathrooms, living room, kitchen, 2 parking spaces, gourmet area, air conditioning for each room excluding bathroom, washing machine , iron, utensils, television, internet, kitchen utensils, refrigerator. To find the standard of rooms, the average of the 30 most profitable rooms was calculated.

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

With the calculated average of the 30 most rented properties in our database for the year 2023, we have:

1. average price database (R$): R$1822.21
2. average rented database (dias): 296.33 ~ 296

Considering the Focus bulletin with the update percentage for the IGP-M as follows:

- 2024: 3.81%
    1. average price database (R$): R$1822.21 * (1 + 3.81/100) = 1891,64
    2. average rented database (dias): 296.33 ~ 296
    3. annual income for apartment = R$ 559924.3
- 2025: 3.99%
    1. average price database (R$): R$1891.64 * (1 + 3.99/100) = 1967.11
    2. average rented database (dias): 296.33 ~ 296
    3. annual income for apartment = R$ 582265.3
- 2026: 4%
    1. average price database (R$): R$ 1967.11 * (1 + 4/100) = 2045.80
    2. average rented database (dias): 296.33 ~ 296
    3. annual income for apartment = R$ 605555.9

Comparative table using random forest machine learning algorithm:

comparing average real x predict**

2023_price = 682.06

2023_predict_price = 681.34

2024_price = 708.04

2024_predict_price = 707.39

2025_price = 736.30

2025_predict_price = 735.59

2026_price = 765.75

2026_predict_price = 764.87

Observation 1:

The average values were different from the values above, previously mentioned, as the basis considered for processing were only types of properties that had at least 30 rentals in the period. In the previous frame was considered the entire base

Observation 2:

The linear regression algorithm was used as a baseline, but its results according to the metrics below were worse than the random forest:

Training data prediction Linear Regression:

Mean Squared Error: 93788.32078540293

Mean Absolute Error: 190.13577357597646

R2 score: 0.5190603875280546

Training data prediction Random Forest:

Mean Squared Error: 2987.4161427966346

Mean Absolute Error: 27.908951979887696

R2 score: 0.9851768357742797

Possible Improvements

For better memory control, we could have broken the pricing file by ad ID, as this is the divisible block we need to run the logic. Furthermore, if this break was made, we could have processed each ID in segregated threads or even in different processes, later we would have had to append the file, without compromising any further processing.

For price prediction we could have used the database, eliminating outliers and adjusted it to run a machine learning algorithm with the Gradient or Random model.

Cross-referencing of VivaReal's property sales file with the possibility of selling part of the apartments built and greater projection of gains not only from rental.

Data detail file with duplicate information, price file with duplicate information, unstructured free text data (important information to organize processing with machine learning) such as amenities and non-standard descriptions of ads. Open text with rich information for filling out rental advertisements makes continuous data processing difficult (natural language processing)

Coding was all carried out as a script in main with the aim of improving understanding of the information. Code can be better adjusted, for example, having a class to handle prices, another to handle details, another with basic rental information, a class to handle opening, closing and general file processing, saving data in a database and retrieve information. This would give the code a more uniform representation and avoid some bad practices of code duplication and isolation of responsibilities.