Seazone Challenge
Data Scientist

This challenge goal is to test your coding skills, logical thinking and analytical capabilities. This is based on the work done in Seazone, with real data. You will be judged both on code structure (variable names, git commits, function abstraction, etc.) and code efficiency (How long does it take to run? How scalable is it? How much memory will your solution require?). All the code you produce will be tested by our technology team, so you must be clear in your README about all the steps needed to produce the intended results.

We also require a report in english covering the problems tackled, the solutions found, and some brief feedback about the challenge. Feel free to expose other possible solutions you considered and possible future improvements to your code. Both the report (in PDF) and your code must be published in a publicly accessible git repository.

We look forward to seeing your results!

1. Data description

Seazone is a Proptech focused on the Short-Stay Vacation Homes market. This market is composed by players such as Guests, Hosts, Real Estate Investors, Constructors and Home Service Providers and we offer the following products and services:

- Property management
- Real Estate Project Development
- Online Travel Agency (reservation marketplace)
- Professional Hosting

As a data-driven company we need to have reliable data and analysis in order to make strategic decisions. In order to do this we built an ETL pipeline based on two main data sources: Airbnb and VivaReal. To feed the pipeline we designed a group of scrapers that acquire the data available online from these websites daily and drop it inside of a data lake. For this challenge we will provide 5 data sets to evaluate your skills in data wrangling, enriching, modeling and also on machine learning.

https://drive.google.com/drive/folders/1ioYOrQobxsGSC-m2V2fJslcALCh2eFnN

Details_Itapema.csv
This dataset contains the data for each airbnb listings based on the listing details such as title, reviews, star rating and description. e.g.
https://www.airbnb.com.br/rooms/70724816509599899 2

Hosts_ids_Itapema.csv
This dataset contains the data available on the host such as the number of reviews and listings. e.g. https://www.airbnb.com.br/users/show/227777128

Mesh_Ids_Data_Itapema.csv
This dataset 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.

Price_AV_Itapema.csv
This dataset contains the price and availability data for a given listing for a given date on an acquisition date.

VivaReal_Itapema.csv
This dataset contains the data for each viva real listings based on the listing details such as price, seller and description. e.g.
https://www.vivareal.com.br/imovel/apartamento-2-quartos-marape-bairros-santos-com-garagem-64m2-venda-RS540000-id-2580913929

2. Data analysis

Itapema is a strategic city for Seazone and we would like to know, based on the data, if we should focus on it or not. In order to make our decision we would like you to tell us the following:

- What is the best property profile to invest in the city?
- Which is the best location in the city in terms of revenue?
- What are the characteristics and reasons for the best revenues in the city?
- 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?
- How much will be the return on investment of this building in the years 2024, 2025 and 2026?

We did not define the terms “best”, “profile”, “location”, “characteristics” on purpose to be able to judge your business intelligence skills.

To answer these questions, feel free to use Excel, Google Sheets, Python, R, Power BI, Datastudio and/or Metabase. Remember to justify your stack choices in your PDF report. **Do not use Jupyter Notebook.**

Remember: Deliver all your work in a git repository! Make sure to make it public, and to add both your code **and** the PDF report to it. Then, you can send the link to the repository via email. Remember also that if we can’t run your code we can’t evaluate it, so make it really clear how to run and make sure it is possible to do so with the files in the repository.