1 BUSINESS PROBLEM

1.1 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.

1.2 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:

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?

2 SOLUTION STRATEGY

2.1 Inputs

Business Questians:

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?

Datasets:

1. Price_AV_Itapema.csv (43020080 rows and 14 colunas)
2. VivaReal_Itapema.csv (17547 rows and 42 columns)
3. Hosts_ids_Itapema.csv.csv (32558 rows and 37 columns)
4. Mesh_Ids_Data_Itapema.csv (2373 rows and 7 columns)

2.2 Outputs

1. A business profile of the best properties cluster, based on their revenue
2. The best localization based on their revenue
3. Designed characteristics of the best properties according to their revenue
4. Designed characteristics and localization according to their ROI
5. The prediction values of the referred years

2.3 Tasks

1. Identify better business metrics for clustering and train a model
2. Use the clustered model and point out the best neighborhoods
3. Use the clustered model and point out the best design features and localization
4. Use the clustered model and point out the best design features location and demand for hosting
5. Create a regression model to forecast demand

2.4 Cicles

1. Create a functional, end-to-end data pipeline (from data collection to model training)
2. Understand the data and clean it (search for inconsistencies) Statistical analyzes first order descriptive
3. Feature Engineering (create variables that model the phenomenon)
4. Exploratory data analysis
5. Define metrics and train the model
6. Analyze metrics
7. Business result analysis and conclusions

3 BUSINESS ASSUMPTIONS

1 Samples with feature available equal to False will be considered how to be generating revenue

2 Future dates and their features will be considered as facts

4 TOOLS AND REQUIREMENTS

1 Language: Python

2 Development: Jupyter Notebook

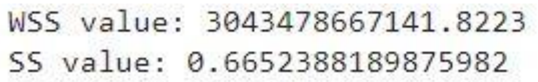
3 Delivery: Public Repository on Github

5 TOP 3 DATA INSIGHTS

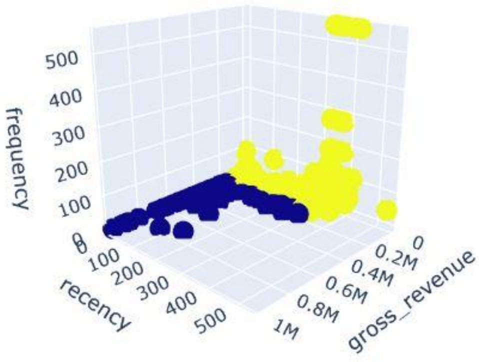
6 CLUSTERING MACHINE LEARNING APPLIED

KMeans

7 CLURSTERING MACHINE LEARNING PERFORMANCE



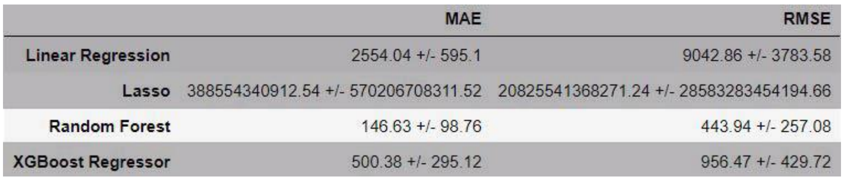
```
WSS value: 3043478667141.8223
SS value: 0.6652388189875982
```



8 REGRESSOR MACHINE LEARNING APPLIED

Random Forest Regressor

9 REGRESSOR MACHINE LEARNING PERFORMANCE



	MAE	RMSE
Linear Regression	2554.04 +/- 595.1	9042.86 +/- 3783.58
Lasso	388554340912.54 +/- 570206708311.52	20825541368271.24 +/- 28583283454194.66
Random Forest	146.63 +/- 98.76	443.94 +/- 257.08
XGBoost Regressor	500.38 +/- 295.12	956.47 +/- 429.72

10 BUSINESS RESULTS

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

Localized at Morretes (average revenue 38% higher than Meia Praia and reasonable demand), with 3 badrooms and 2 bathrooms.

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

Ilhota (considering the average just revenue, without considering the demand)

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

Distance from the beach, Neighborhood, Bedrooms quantities

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?**

Answer

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

Answer

11 NEXT STEPS

- Better understand the reliability of working with future events, as data from 2023 are contained in the dataset
- Effect size studies and A/B testing on defined clusters, oriented towards financial metrics such as gross revenue

- Model deployment