# Report: Seazone Challenge Jr DS

## 1. Introduction

This report is intended to detail the methods and tools used to resolve the '*Seazone Challenge Junior Data Scientist*'. The step by step used in the data exploration part will be discussed, possible errors and solutions found in the data, exploratory analysis, survey of insights and possibilities for future solutions and implementations.

## 2. Case

Seazone is a company that sells stays of different types of properties, which other travel agencies use in addition to their own website.. For this challenge, two data tables were made available, one with details of the listings such as location, furniture, capacity, property type and a second table with listing details such as values, listing date and lease, property availability.

## 3. Solution planning

- What will be done:
  - Visualization and initial exploration of the data, to understand them using python and Google Sheets to visualize the tables.
  - Cleaning the data, organizing and renaming columns, describing the types, dimension, checking missing data and performing statistical description, creating business hypotheses creating new variables to explore these further.
  - Exploration of the data, using the Power BI tool, some dashboards with interactive filters will be created. Some business hypotheses and validation of them will be created.
  - Using dashboards built using the data already treated to answer the questions of the challenge.
- Tools used:
  - Python 3.9
  - Pycharm 2022.1
  - Google Sheets
  - Power BI Desktop
  - Power Bi Editor Power Query

## 4. Data Description

- What was done**:**
    - To explore the data, an initial visualization was made using the Google Sheets tool. Through an IDE and using the python language and its own libraries, the data was loaded, the columns were renamed ‘listing’ from the table “listings-challenge” and the column “Código” from the table “daily_revenue”, both for the name “id”, standardizing them as key column, to facilitate merging between tables. Thus, the tables were merged to work with a single table with data from both tables.
    - In this step, we check the dimensions, types of columns and the existence of NaNs, to perform possible treatment if necessary.
    - Renamed all columns to snake_case pattern, to facilitate data manipulation using python.
    - At the end, a new data table was generated and saved with the name of “df_listings_revenue_1”
- Problems found**:**
    - Some features were identified when loading data in the IDE with wrong type and will have their types changed in the next step.
    - Several NaNs were found, will be investigated and dealt with in the next step.

The development of this step is in the file ‘data_description.py’.

- Tools used:
    - Ide Pycharm with Python: We chose to use python because it is a language with several libraries that facilitate the investigation and manipulation of data in a quick and practical way.
    - Google Sheets: Tool that made it possible to open, visualize and manipulate data in an online and intuitive way.

## 5. Data Cleaning

- What was done**:**
    - First, an investigation of the wrong data found during the data exploration in the previous step was carried out. Some columns that indicate numbers such as furniture, bathrooms for example, were identified as Object because they had entries replicating the title instead of indicating the number. The treatment chosen here was to replace these texts with 0, because important information to answer the questions, such as revenue, occupancy and dates were ok, it was the way to continue with the data in the base and deal with these errors.
    - The address variable, there were some addresses that were not filled in, so the treatment done here was to replace these with 'NONE'. Because the rest of the data from these records were ok.
    - Several features that indicated quantities were not filled in, consisting of NaN, for these and in order to maintain the data as they had important registration information, we decided to change them to zero, taking care that if they are used, knowing that they have untrue data.
    - The feature 'occupancy' had some records with fill '2' however, as it is a feature that indicates 1 or 0, we assumed a typing error and the 2 was changed to 1, because when investigating the table, these records had been invoiced, so they were supposed to be 1.
    - The feature 'creation_date' phas several records such as 'NaT', as this is only filled when the property is leased, so it will only be used in this situation for analysis, then we will leave it as 'NaT', when the property is 'occupancy' =1 and 'blocked' = 0
    - It was identified that the advertisement "TST001" had all the records unfilled, so it was decided to exclude these records as they will not be useful in our analysis.
    - After this first treatment, we created two new tables, one with all the data from the emerge revenue and listings table that was named as "df_listings_revenue_total.csv" and another with only the listings when occupancy and not blocked which was named as "df_listings_revenue_occupancy"

These treatments were performed in python, and the file with its development is in the *"data_cleaning.py"* file.

- Problems found:
    - In addition to the errors found earlier and dealt with in this step, there was an inconsistency in some date records, where the 'creation_date' should always be smaller than the 'date', but some records do not obey this logic. At first, no treatment will be done in relation to this, but this error will be taken into account when analyzing where this difference between the dates will be used.
- Tools used:
    - Ide Pycharm and Python: We continue to use python for data processing for reasons already exposed.

**6. Statistical analysis of the data**

- What was done**:**
  - First, the data was separated into numerical and categorical, so that we could make a descriptive and statistical analysis of the data after the first treatments, aiming to have a broader view of the data and analyze more broadly any possible inconsistency or problem based on these analysis.
  - The separation into categorical and numeric, was done in order to simplify, it was separated with all the data that until now classified as Int and float would be analyzed as numeric and the data of other types as categorical.
  - Numerical data were analyzed for the size of its range, maximum and minimum values, mean, median, standard deviation, skew and kurtosis.
  - At first, no abnormalities were found in this statistical study of numerical variables. The features that indicate price, there was a strong concentration on the number zero, expected since in this analysis the data were not filtered with the days where the property was empty, and due to this large number of days that remain without being logged, either because are available or out of rental seasons, this concentration of null values occurs in the distribution, these variations will be analyzed in the stage that will be carried out by the EDA.

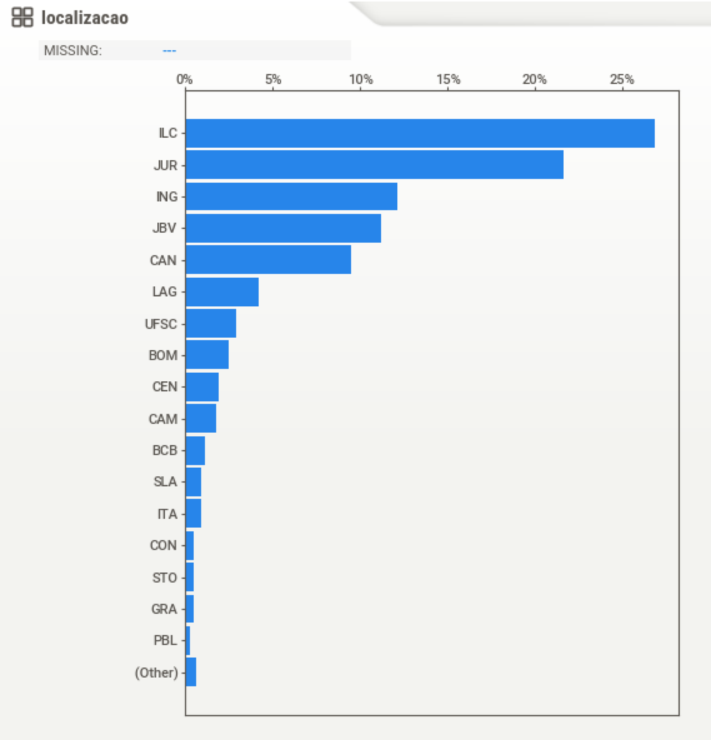
- Problems found:
  - This analysis was a little hampered by the amount of '0' in some features such as 'revenue' and 'creation_date' which are 0 when the property is empty. However, as we separate it into two different tables, one with the total data and one just filtering the properties when leased to facilitate future analysis, this will not be a problem.
- Tools used:
  - Ide Pycharm and Python: The calculation libraries available in python make this process easy and fast, hence the choice of this language and tools.

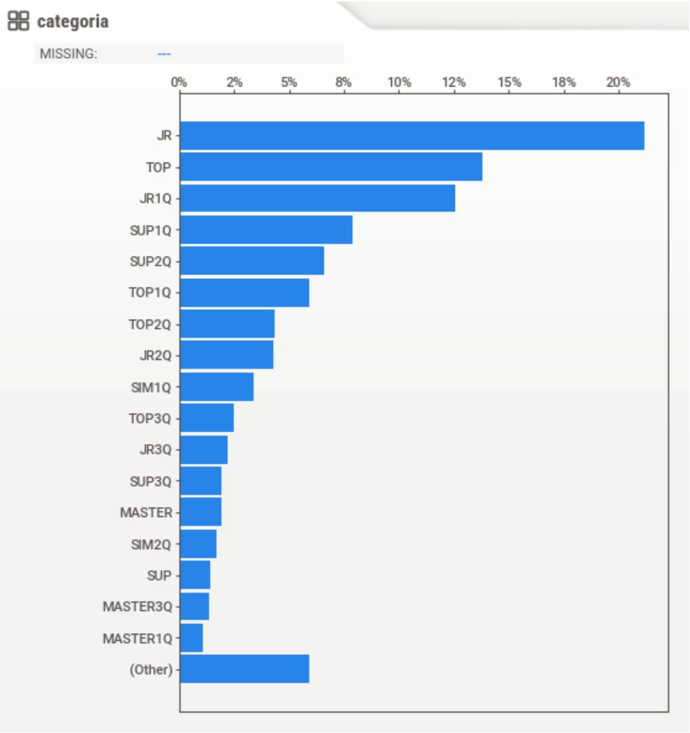
The generated table can be found in the 'df_num_statistics.csv' file, inside the 'statistics' folder that is in the 'python_project' folder in the github repository of this project.

## 7. **EDA(Exploratory Data Analysis)**

### 7.1 Uni-variate Analysis

- What was done**:**
  - This analysis is carried out in an isolated study of each variable, in order to understand the variable in isolation, regarding its distribution. To speed up the process, the sweetviz module was used, which generates a simplified report of each feature in a dependent way so that we can analyze its distribution, the existence of conflicting data.
  - After analyzing each feature, no more abnormality was noticed in the data that would prevent the development of the project from proceeding.
  - When analyzing the categorical variables, the quantity of each category was analyzed, to visualize the balance of data and quantity of data that we had for each category. Here, it can be seen that in the 'type' column we have a low number of the 'type' 'House', with only 5% of this type in the database. In the column 'location' we also have a high concentration of some localities, as can be seen in the graph below. The same occurs in the 'category' column where some categories like TOP, JR and JR1Q have a large number of data when compared to other categories.





- Tools used:
  - Python e Sweetvis_report: The Sweetviz_Report application is a function created in python, it was used to facilitate plotting and speed up the process, as it automatically plots a visualization and statistical data of each variable independently.

These generated graphics were saved in the file 'SWEETVIZ_REPORT.html' inside the folder “python_project” present in the github of this repository.

### 7.2 Bi variate Analysis

This step was made by exploring and visualizing, crossing features. After the initial analyzes and explorations and looking at the end of the report to resolve the issues raised, a dashboard was developed using the Power BI tool, separated into 4 main tabs, "Billing", "Reservations", "Location" and 'New Year' . These dashboards have filters with the most important features for the business and graphs crossing the data, where different business questions can be answered. The "Reservation" dashboard used data from the dataframe treated with all the data, 'df_listings_revenue_total.csv', the other two were used to filter only the properties that obtained billing, 'df_listings_revenue_occupancy.csv'.

This dashboard was developed using Microsoft's Power BI, this choice was made because it is a very intuitive program and with a good range of features for building and visualizing graphs.

The data used was the data already processed in python in the previous phases of the project development using pycharm.

Several business hypotheses were raised and answered using these visualizations. These business hypotheses were made in order to increase the knowledge of the business through the available data and generate Insghts for the company.

At the end we separate 3 business insights based on these business hypotheses, these are answered in Readme on github.

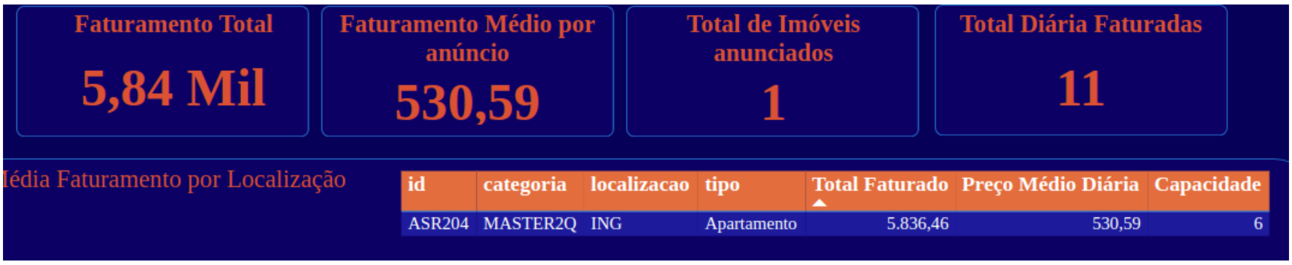
All hypotheses generated and responses obtained with their respective graphs are in the separate report named "hypoteses_negocio" and available in the Git Hub repository

The Power BI project is available in '.pbix' format in the '.df' folder in this project's github repository. Along with pdf files showing the 4 tabs of the developed dashboard in "seazone_challange_dashboard.pdf"

## 8. Questions:

***1. What is the expected price and revenue for a listing tagged as JUR MASTER 2Q in march?***

- In the available data, no advertisement of this type with JUR location was found.
- Thus, to estimate a price for this listing, the following strategy was used:
    - First, a filtering was done using the built dashboard 'Location', selecting properties with category 'MASTER2Q'. However, only 1 property of this type was found, which has already been announced and leased 11 nights in March. With the value of 530,59 reais. Located in 'ING', as we can see in the image below. So, taking the average revenue from properties in 'ING' in March is 279 and 'JUR' has an average revenue from their ads of 351. Using a simple rule of 3, we have R$ 666,00.



Faturamento Total
5,84 Mil

Faturamento Médio por anúncio
530,59

Total de Imóveis anunciados
1

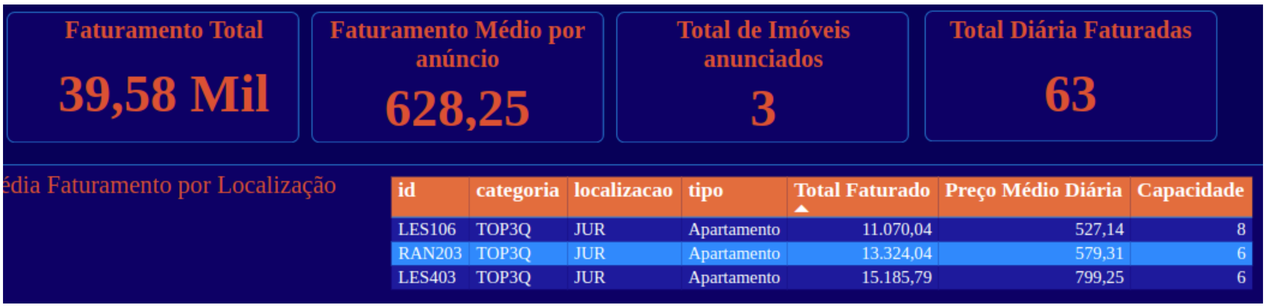
Total Diária Faturadas
11

lédia Faturamento por Localização

| id | categoria | localizacao | tipo | Total Faturado | Preço Médio Diária | Capacidade |
|---|---|---|---|---|---|---|
| ASR204 | MASTER2Q | ING | Apartamento | 5.836,46 | 530,59 | 6 |

*Filter result March month and category MASTER2Q*

- 
    - Analyzing the average revenue of 'category' with values closer to 'MASTER2Q', ads were found in 'JUR' of type 'TOP3Q' which, on average, has a value of 534 for the month of March, while 'MAster2Q' has 531. Thus, carrying out the search only for 'JUR' of the 'TOP3Q' category in the month of March, we have an average value of 628.25 reais. Again making a simple rule of 3, with the average values, the value of 624,72 reais was reached.
    -



Faturamento Total
39,58 Mil

Faturamento Médio por anúncio
628,25

Total de Imóveis anunciados
3

Total Diária Faturadas
63

édia Faturamento por Localização

| id | categoria | localizacao | tipo | Total Faturado | Preço Médio Diária | Capacidade |
|---|---|---|---|---|---|---|
| LES106 | TOP3Q | JUR | Apartamento | 11.070,04 | 527,14 | 8 |
| RAN203 | TOP3Q | JUR | Apartamento | 13.324,04 | 579,31 | 6 |
| LES403 | TOP3Q | JUR | Apartamento | 15.185,79 | 799,25 | 6 |

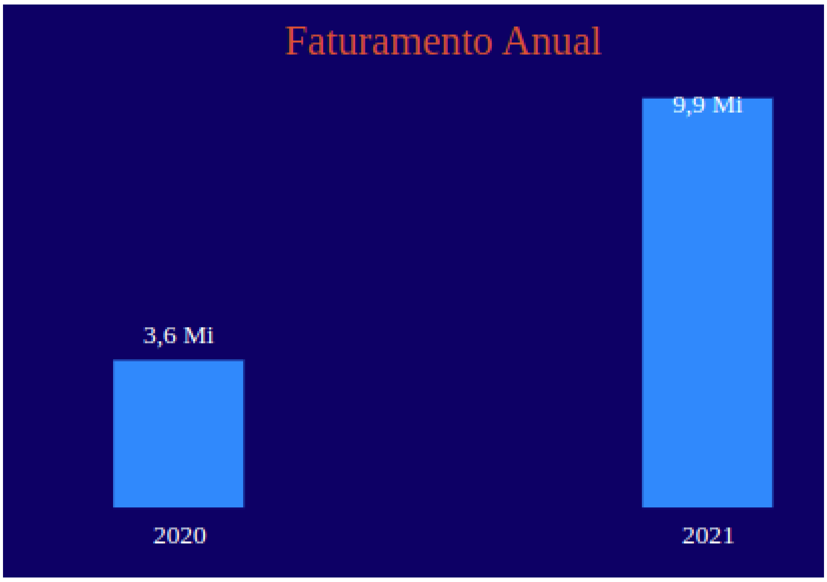
Filter result month March category TOP3Q localization JUR.

So we can estimate that a good price to be charged for a JUR MASTER2Q property in March is 645 reais, which is the average value of these two estimates, which were very close.

***RESPOSTA: R$ 645,00***

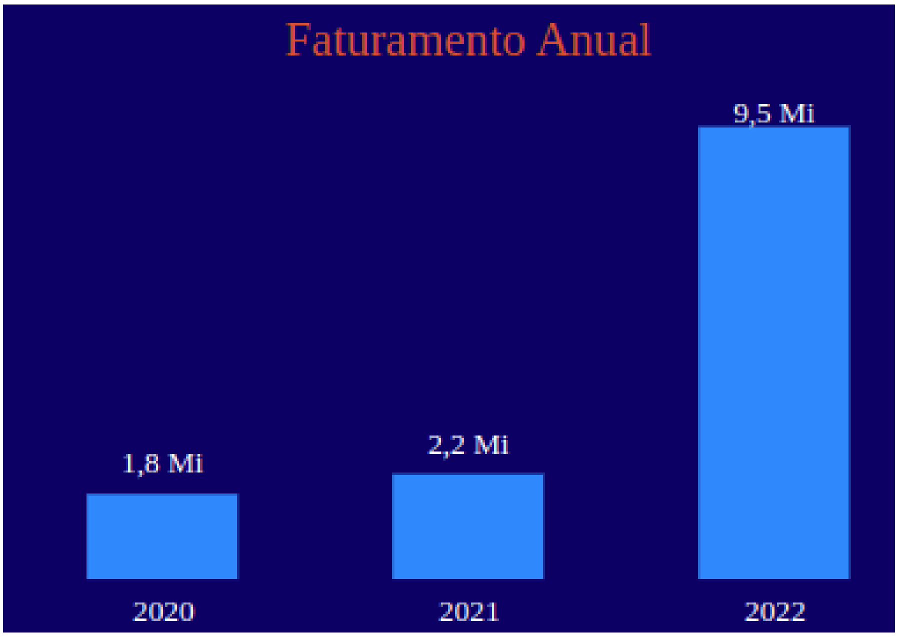
### *2. What is Seazone's expected revenue for 2022? Why?*

- Analyzing the revenue data, from the two complete years we have, 2020 had a revenue of R$ 3,6 million  and 2021 R$ 9,9 million. A 275% growth in revenue. In 2022, with the data up to March, it already has a turnover of R$ 10,2 million. Taking into account only these growth estimates, we could expect a turnover of R$ 27,22 million.



*Total Revenue years 2020 and 2021*

- When we analyze revenue month by month. Analyzing the revenue as far as we have the 2022 data for comparison, we have revenue in 2020 in the first 3 months of 1,79 million, in 2021 of 2.2 million and in 2022 of 9,5 million. A growth of 330% in these first 3 months. On top of this performance, a turnover of 40,8 million would be expected. However, if we are to analyze not only the data, but also the real scenario where we had a pandemic that caused these numbers to be harmed at the beginning of the year 2021. This can be seen when comparing the monthly billing graphs for 2020 and 2021, where 2020 had a close performance in January, February with the last months November and December, which are the highest revenue months due to the season.



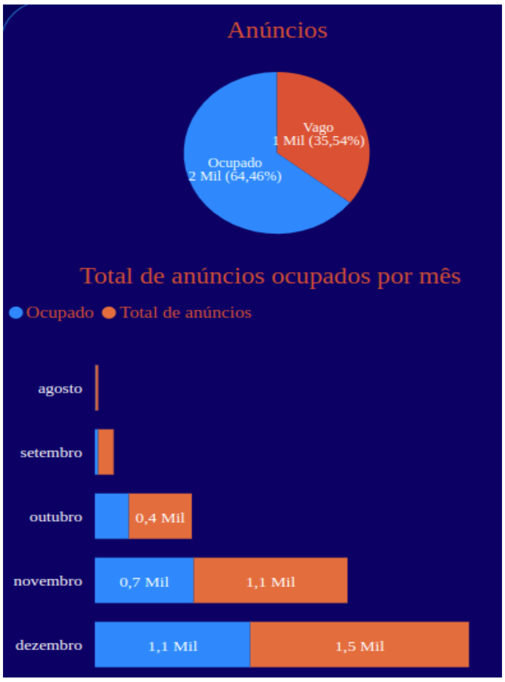
*Revenue January, February and March*

- Making a projection of 2022 with a scenario less affected by the pandemic, a turnover similar in the last month to what it was in these first months, considering seasonality, and an average similar to that of March for the low season months. We would have another R$ 9,5 million  in recent months and an average of R$ 1,7 million in revenue from April to September. Resulting in R$ 29,2 million in revenue.
- Thus, taking into account expectations on top of the most realistic numbers, and an already growing number of ads. We can estimate a revenue of around R$ 30 million  after analyzing all these numbers and expectations, based on the data we have and taking into account the crisis caused by the pandemic.

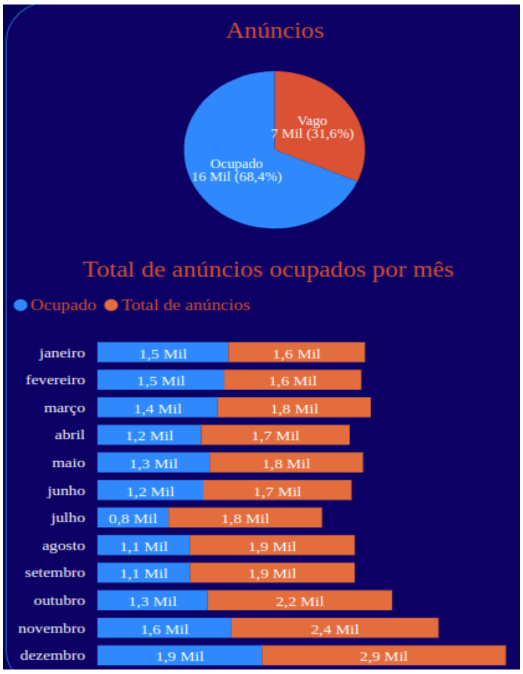
***RESPOSTA:  R$ 30 milhões***

### *3. How many reservations should we expect to sell per day? Why?*¶

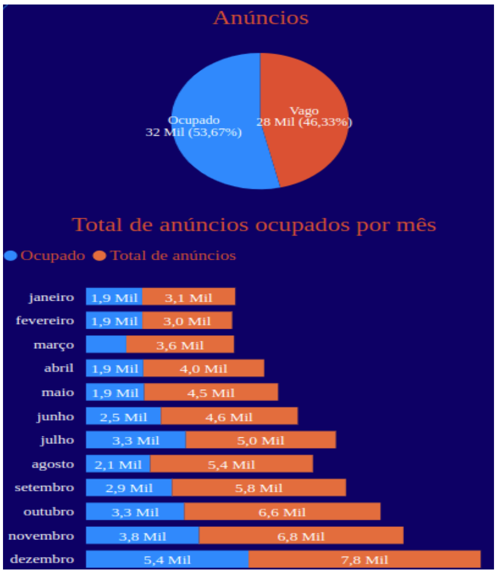
- Analyze year by year:
  - At the end of 2019, Seazone had 51 properties with 1.500 listings in December and an average occupancy rate of 64%.
  - At the end of 2020, it closed December with 99 properties advertised and 2.900 advertisements, with an average occupancy of 68%.
  - At the end of 2021, it closed December with 283 proprieties and 7.800 advertisements, with an average occupancy of 53,67%.
  - In 2022, until March there were 317 proprieties and 9,8 thousand listings, with an average occupancy rate of 76,5%.



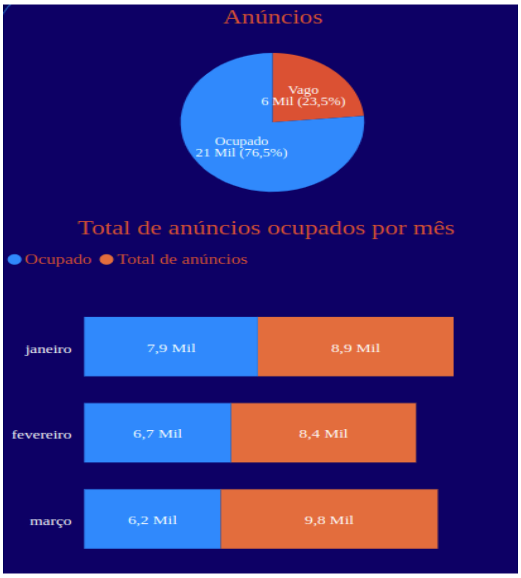
*Ano de 2019*



*Ano de 2020*

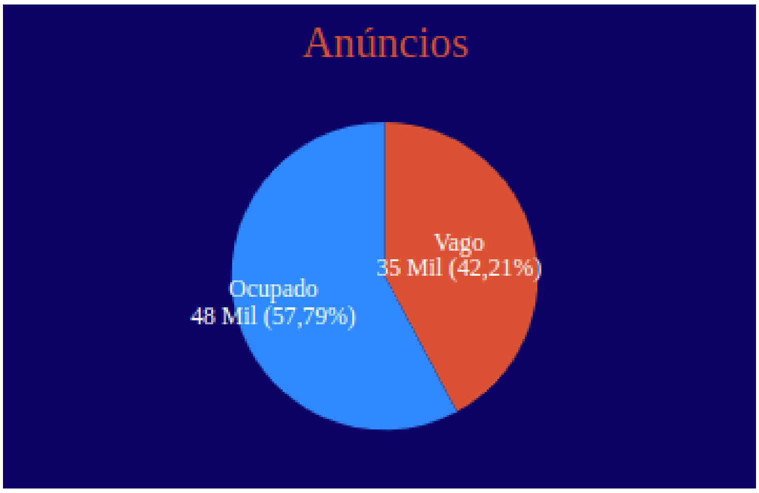


*Ano de 2020*



*Ano de 2022 meses Janeiro, Fevereiro e Março*

  - There was a 93% growth from 2019 to 2020 and a 168% growth in the number of ads from 2020 to 2021. Considering that there has already been a growth in the first 3 months of 2022 of 25,64% in the number of listings.
- The average occupancy considering the years 2020 and 2021, for which we have data for the entire year, was 57,8%. Considering that until March the number of properties announced was 317 properties. Thus, having an average occupancy of 57,8% of occupancy of these properties, we can expect an average of 183 properties reserved per day.



*Ano de 2020 e 2021*

- This number tends to grow considering the growth in the number of properties in its base that Seazone has had over time.

***<u>RESPOSTA:  An average of 183 reservations per day is expected</u>***

### *4. At what time of the year should we expect to have sold 10% of our new year'snights? And 50%? And 80%?*

To answer this question, the dashboard on the 'New Year' tab was used, where we have a table with the number of days of reservation, id and 'creation_date'. As we have already seen, we have some reservations with 'creation_date' greater than 'date', so this table was used to check the ids and remove them from the filtering.

To obtain the % of reservations sold, a waterfall chart was made, where the 'creation_date' was plotted on the x axis and the % of the ids count for each date, so we were able to obtain on which dates the amount % of vacancies for each new year date.

The filters were set on the 31st of the 12th month, changing the year and filtering the ids with negative reserve days.

- In 2019, 51 properties were announced in Sazone on the night of 31/12.
  - On 06/11/2019, 10% of the vacancies had already been sold, 55 days in advance.
  - On 06/12/2019, 50% of the vacancies had already been sold, 25 days in advance
  - On 17/12/2019, 80% of the vacancies had already been sold, 14 days in advance.
- In 2020, 99 properties were announced in Sazone on the night of 31/12.
  - On 27/10/2020, 10% of vacancies had already been sold, 71 days in advance.
  - On 13/12/2020, 50% of the vacancies had already been sold, 18 days in advance
  - On 26/12/2019, 80% of the vacancies had already been sold, 5 days in advance.
- In 2021, 283 properties were announced in Sazone on 31/12.
  - On 05/10/2020, 10% of the vacancies had already been sold, 85 days in advance.
  - On12/12/2020  50% of the vacancies had already been sold, 19 days in advance.
  - On 24/12/2019 80% of the vacancies had already been sold,, 7 days in advance.

- Based on the data, an average of days will be estimated for each percentage, in order to obtain an estimate of occupancy for each % requested.
  - 10% of vacancies will be filled 70 days in advance
  - 50% of vacancies will be filled 20 days in advance
  - 80% of vacancies will be filled 8 days in advance

***RESPOSTA:  of the year should we expect to have sold 10% : 70 days in advance,***

***50% : 20 days in advance and 80% : 8 days in advance***

### *4.1. How can this information be useful for pricing our listings?*

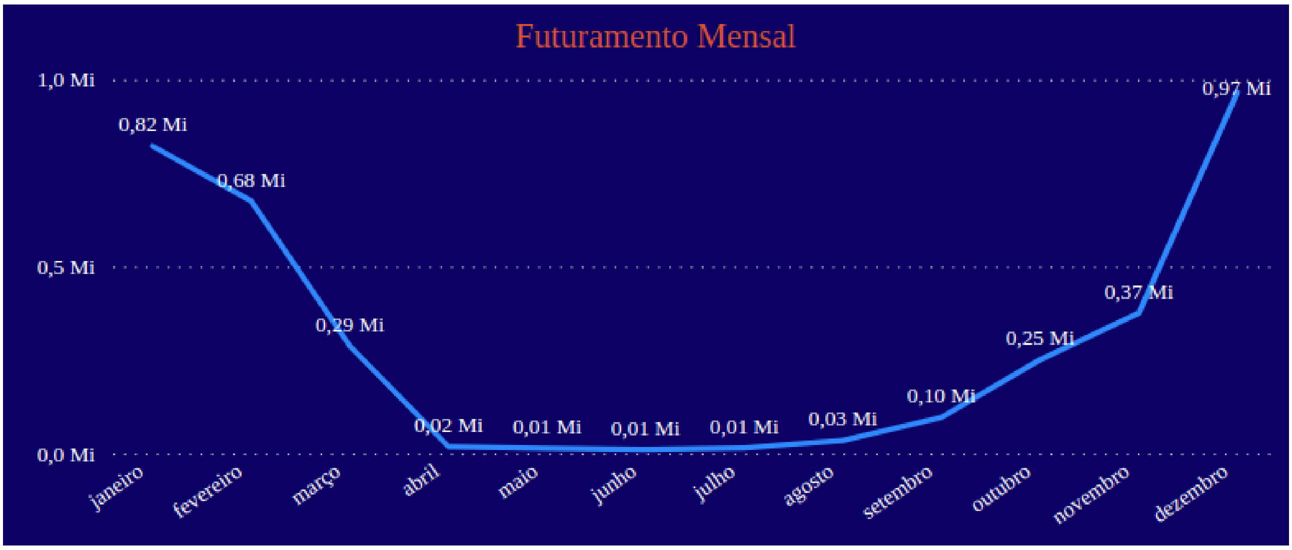
This information is important for pricing and estimating daily packages, knowing how soon there will be fewer offers for properties for this date, it makes it easier to assemble the package strategy for the new year and for dates around the new year.

***Questão extra:***

***5. On the impact of the COVID-19 pandemic:***

***- Can we estimate Seazone's revenue loss due to the pandemic? How?***

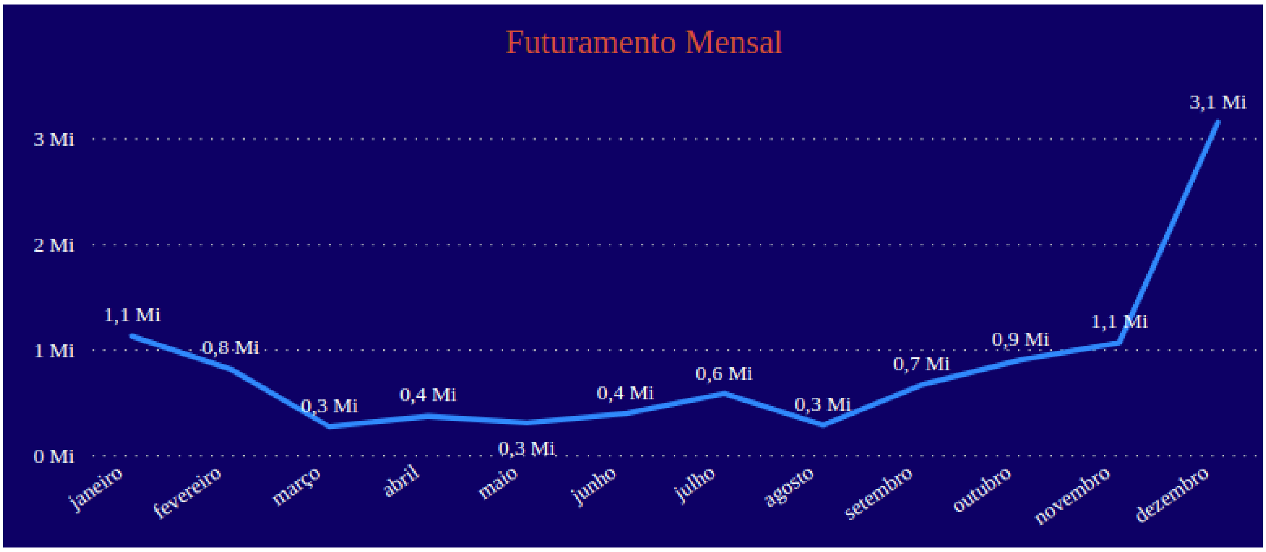
When plotting a month-by-month billing graph for the year 2020, we noticed a sharp drop from March, the month that the pandemic restrictions began here in Brazil.



*Plot 2020, revenue each month*

Even with the growth in the number of properties advertised each year by seazone, this revenue grew modestly at the end of the year, still suffering from the pandemic.

We can see in the graph below that in 2021, it continued to suffer from the pandemic, only having a recovery with its weakening at the end of the year. Finally having a strong growth in billing, managing to bill about 3x more than in the previous year. This same time of year end, which is usually the highest billing season.



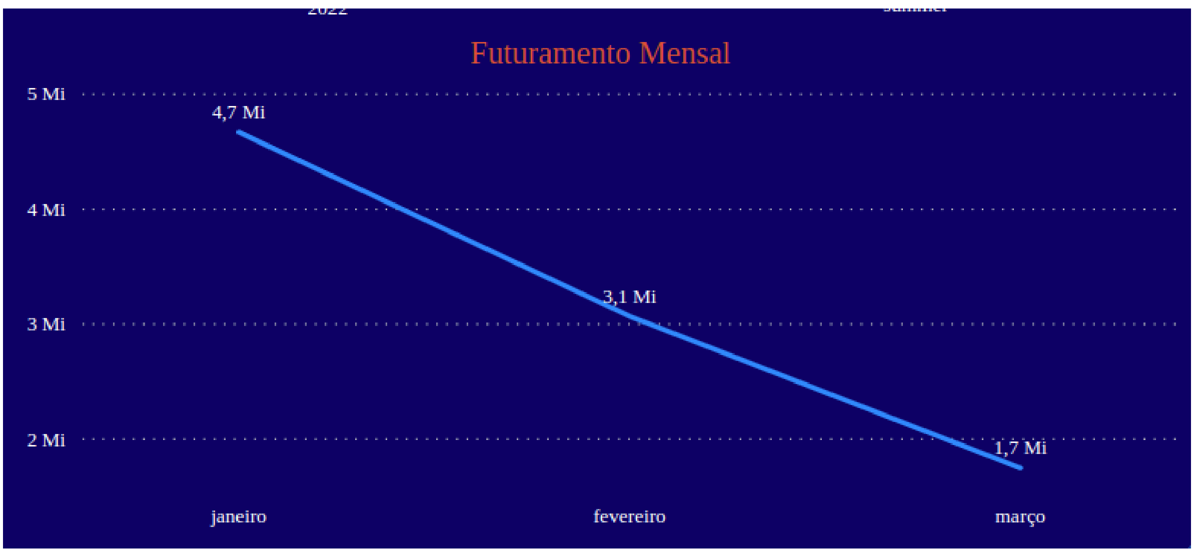
Plot 2021, revenue each month

***<u>RESPOSTA: Analyzing the plotted data, we can estimate strong losses in billing due to the pandemic.</u>***

***- Has the industry recovered?***

It is possible to notice a recovery at the end of 2021, confining this recovery to the first months of 2022. Following this trend, we can say that we have a recovery in progress and we can expect a normalization for the next year-end season in the sector.

Even noticing a drop, this due to the end of the season, where we always expected a drop after the months of January and February, it was no longer as sharp as in previous years.



*Plot 2022, revenue January, February and March*

***<u>RESPOSTA: It is in the process of full recovery and should be normalized by the next high season.</u>***