Seazone Challenge

1. Introduction

The Seazone Challenge aims to answer some questions about the business in Florianópolis. The company has a range of properties in different neighborhoods that can be classified from simple to more luxurious hotels, apartments or houses.

The following research will manipulate the data to identify patterns and insights to solve the problems presented. Next, machine learning models were used to predict some results and answer the questions.

2. Preparing the Data

To answer the questions and solve problems, it is needed to know the data and prepare it. Thus, we will have accuracy in the data, which leads to accurate insights. Without data preparation, it is possible to have wrong insights due to junk data.

In the Seazone Challenge there are two different tables that complement each other: the listing and the daily revenue table.

Firstly, we are going to consider the listing table. We can see that there is some missing information about pillows, cleaning fee, beds and among others. However, it is not possible to just delete these missing values or make an average to replace them, because it directly affects the customer's decision whether to rent or not. Furthermore, this is a table of all the places that Seazone has a contract with and it can not be excluded. This missing information will not affect the data analysis and it can be collected in another moment.

In the revenue table, the “creation_date” column of a given row should be always lesser or equal to the “date” column of the same row. Therefore, the rows with the “creation_date” higher to the “date” will be excluded. After this procedure, the revenue table does not have any null values and is ready to go.

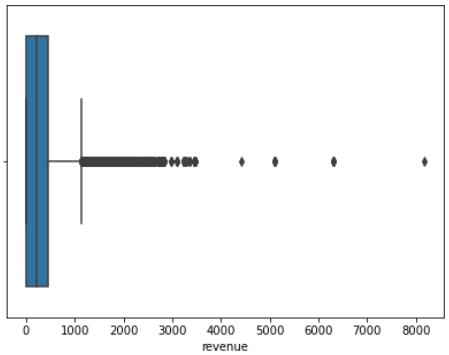
The “date” and “creation_date” columns were converted to a date time variable so it can help manipulate data later.

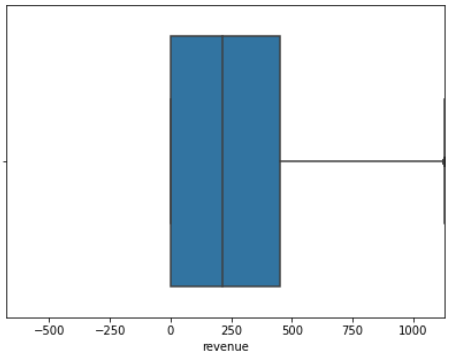
a. Treatment of Outliers

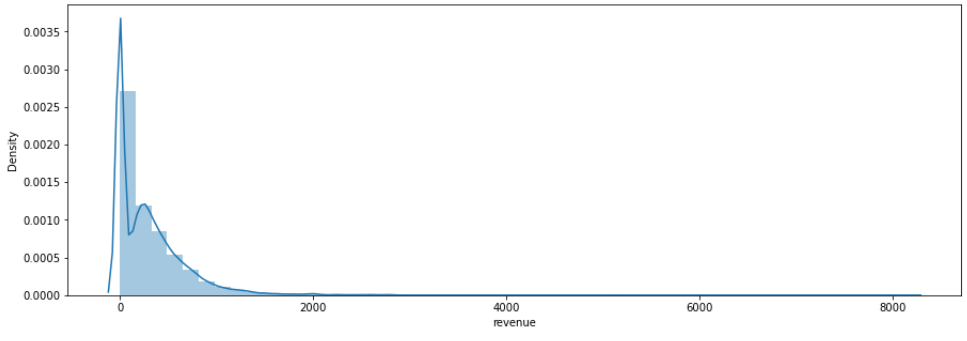
The Outliers must be analyzed because they can influence our machine learning model and predict wrong results.

Firstly, a heat map was created to see the correlation of each feature in the revenue table. It is possible to notice that the last_offered_price has a strong correlation with the revenue.

Then, functions were created for outlier analysis.It will be used as rule values below Q1 - 1.5x Amplitude and values above Q3 + 1.5x Amplitude (Amplitude = Q3 - Q1).





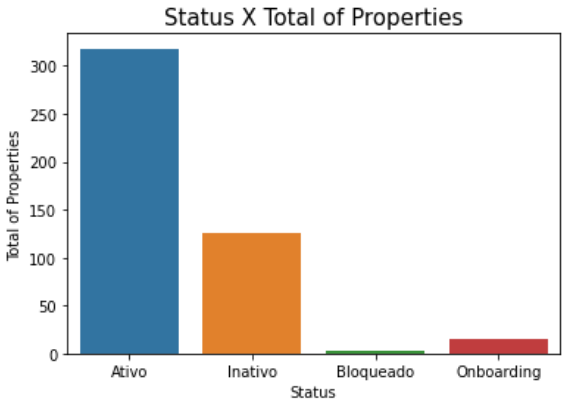


It is clear to see in the graph that the dataframe contains some outliers. However the revenue of these outliers is estimated at R$3.544.412,62 in the period given and, for this reason, will not be considered as an outlier.

Furthermore, some outliers could be generated by the properties classified as MASTER, because it is a more luxurious categorie with higher prices. Thus, these properties will not be considered outliers because even representing just 4.33% of the properties available, they represent 8.7% of the total revenue in all the periods given, that is equivalent to R$2.110.253,15. Therefore, they will not be deleted from the data.

3. Exploratory Data Analysis

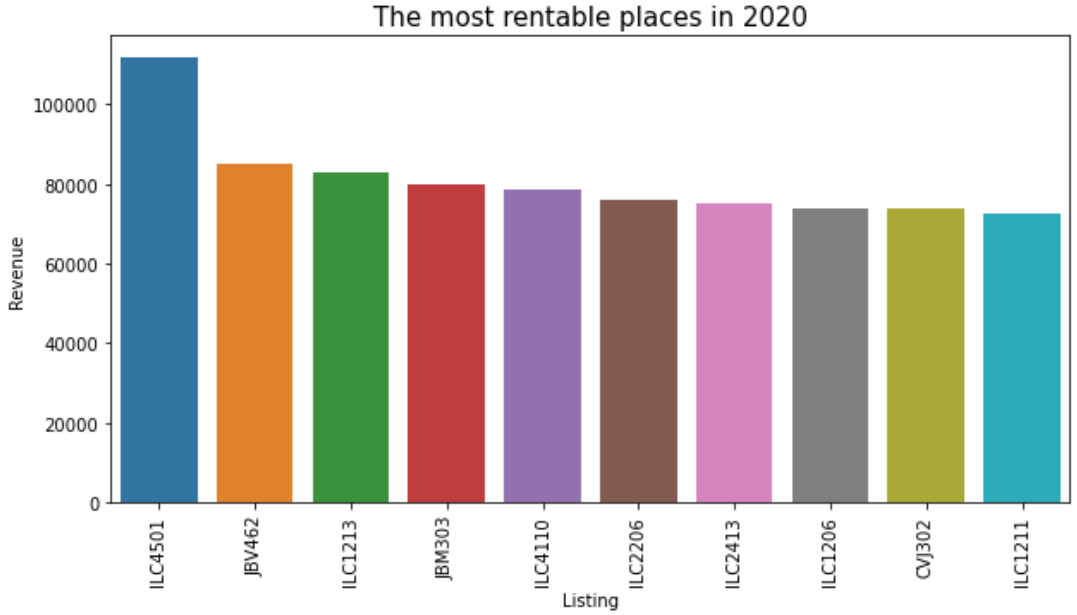
With the data prepared, we can do the exploratory data analysis to expose trends, patterns and relationships that are not readily apparent.

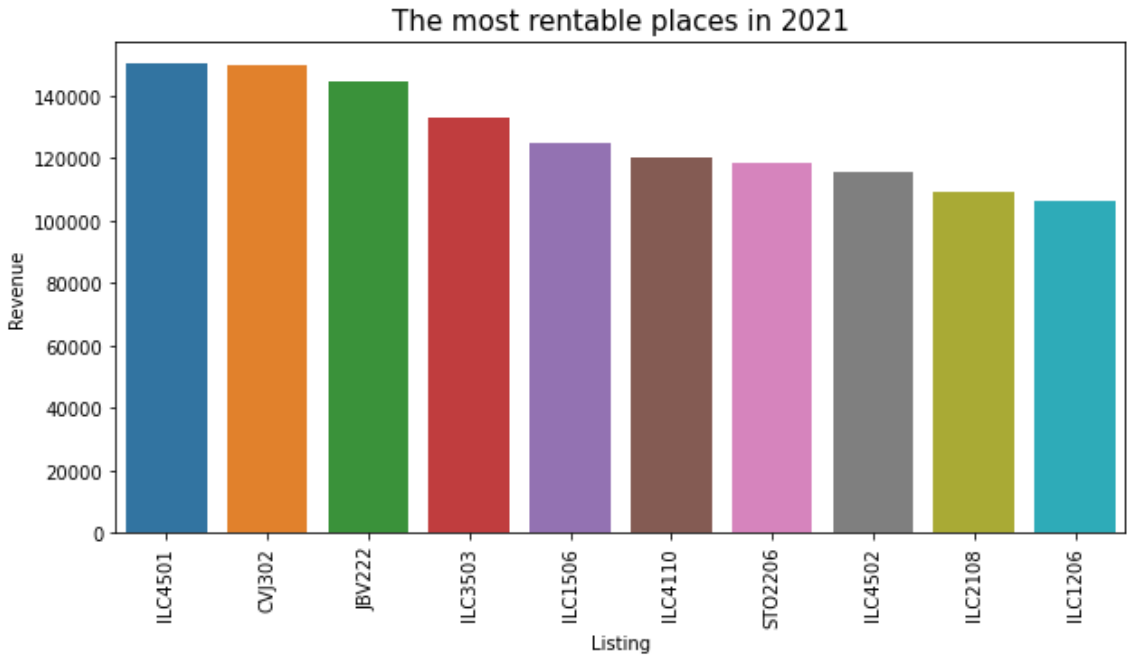


There are 462 properties, in which 318 have the status as active, 128 inactive, 15 onboarding and 3 blocked. However, it is possible to notice that there are too many inactive properties and the analysis could not identify a pattern between location, categories or type. For this reason the Seazone team should enter in contact with the landlords to know if there is a common reason between them and if there are any chances to be active again.

The properties mentioned above 264 are apartments, 172 are inside a Hotel and 26 are houses and the majority are located in JUR and ILC.

When it comes to the total revenue of 2020 and 2021, they were R$3.536.717,09 and R$9.846.212,90 respectively, that is the revenue in 2021 was 178.0% higher than 2020. There is an extreme difference between those two years, possibly due to the global pandemic and consequently the lockdown. Although, in both years it is clear to see that the top 10 most rentable properties are those starting with the code “ILC”, that is a hotel in JUR.





4. **Answering the Questions, Problems Tackled, Solutions and Improvements**.

The first step to answer the question is to adjust the features to make the work of the future machine learning model easier, this is called Encoding. However, this step will not be needed because the features are already prepared.

a. Machine Learning Model

To all the questions that we are going to use machine learning models to answer, there are the same basic principles, which will be explained next.

The parameter R^2 will be used, which tells how good our model is: the closer to 100%, the better. We are also going to calculate the Mean Square Error, which will show how wrong our model is: the smaller the error, the better.

Three machine learning models will be used: Random Forest, Linear Regression or Extra Tree.

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

Filtering the listing table, it is not possible to find any property within the category MASTER2Q located in JUR. However, there are different properties classified as MASTER in JUR and for research perspectives we are going to use the MASTER3Q in JUR, code CVJ302, because it has more data to work with.

Within the models used, the Extra Trees Model has more accuracy and, therefore, the price and revenue were calculated with this method. The expected price and revenue for a listing tagged as JUR MASTER 3Q in march is R$662,71 and R$753,28 per day of rent respectively.

It is possible to notice that the revenue is higher than the last offered price, this happens because in the dataset there are some rows with this configuration. I would rather not delete these data because there could be some extra fees that were not mentioned. Example: When a property is suitable for 3 people but it is possible to pay an extra fee to let more people in.

Besides that, this model can be improved by creating a Deploy of the project in Streamlit, which can be an interactive tool that can predict any time of the year of any property just entering the code and the date.

c. What is Seazone's expected revenue for 2022? Why?

To find out what the expected revenue for 2022 is, a table with the monthly revenue from august 2019 to february 2022 was created. The january and february of 2022 are closed months. Therefore, they do not need to be predicted.

This data was used to train our machine learning models and the random forest method was the one with more accuracy. The expected revenue for 2022 is around R$40.000.00,00.

However, the accuracy of this model is too low and the error is high, so this cannot be considered a faithful method. To answer this question properly we need more historical data.

d. How many reservations should we expect to sell per day? Why?

A single reservation can be composed of many consecutive nights. For this reason, to answer this question, the 'creation_date' will be used, which is the date that the reservation was made.

To know how many reservations were done per day, the 'creation_date' must be equal to the 'creation_date' of the following day and different to the past day. Thus, a table of each day that any reservations were made in 2021 was created and used as a basis to train our machine learning model.

	month	reservation_per_day
0	1.0	11.0
1	2.0	10.0
2	3.0	6.0
3	4.0	12.0
4	5.0	11.0
5	6.0	12.0
6	7.0	17.0
7	8.0	12.0
8	9.0	15.0
9	10.0	16.0
10	11.0	30.0
11	12.0	32.0

The Extra Tree model was the more accurate and with the lower error. We use this model to predict how many reservations per day would be done in 2022. However, the expected reservation for 2022 is not compatible with the seasonality because the results were almost the same for each month. For this reason, we are going to disconsiderate this evaluation and to calculate we are going to use the 2021 original data as a base.

The expected reservations per day can be seen in the table beside. It is not possible to say a unique number of reservations for the whole year because due to the seasonality, that is, December is the month of higher selling with an average of 32 reservations per day and February is the month of lowest selling with an average of 6 reservations per day.

e. At what time of the year should we expect to have sold 10% of our new year's nights? And 50%? And 80%? How can this information be useful for pricing our listings?

When we filter the date of December 31 in the revenue table regardless of the year, we have little data available to create a machine learning model. For this reason, I found it more suitable to work with the historical data of 2021, because this question is about customer behavior and I think that this comportament will repeat over the following years regardless of the amount of the reservation.

Nowadays, there are 318 active properties available to rent, but 38 landlords block their properties for new year's nights, resulting in 280 properties. Thus, 10% of the new year's nights is equivalent to 28 properties rentented, 50% is equivalent to 140 and 80% is equivalent to 224 properties.

This way, we can find the results easily filtering the revenue table.

- 10% of the new year's nights should be sold by October.
- 50% of the new year's nights should be sold by December.
- 80% of the new year's nights should be sold by the end of December.

This measure can help to price the properties according to the laws of supply and demand, that is, the scarcer the properties are becoming, the more expensive.

f. On the impact of the COVID-19 pandemic:

Due to the short amount of time I was not able to answer the following questions properly, but I will describe a guideline so they can be solved.

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

The Seazone's revenue loss can be estimated by creating a machine learning model with the previous data of 2018 and 2019 to predict the 2020 revenue. Therefore, the forecast revenue subtracted from the actual revenue will give the revenue loss.

ii. **Has the industry recovered? If Yes, when can we state that we came back from pre-pandemic levels of sales/revenue? If No, when do you expect this recovery to happen?**

Yes, it is possible to say the industry recovered because the revenue of 2021 was 178% higher than 2020. However, the data available for analysis begin in the second semester of 2019 so it is not possible to measure the pre-pandemic levels of sales/revenue.

5. Feedback

In my personal opinion, it was very challenging to create the codes within the time given. It was my first machine learning model that I created totally by myself and without supervision. I believe that I have a lot to contribute, but also, so much to learn. I hope that my passion and my willingness to always learn more and improve will help me to pass the next step of the selective process.

I'm looking forward to the interview and thank you for the opportunity!