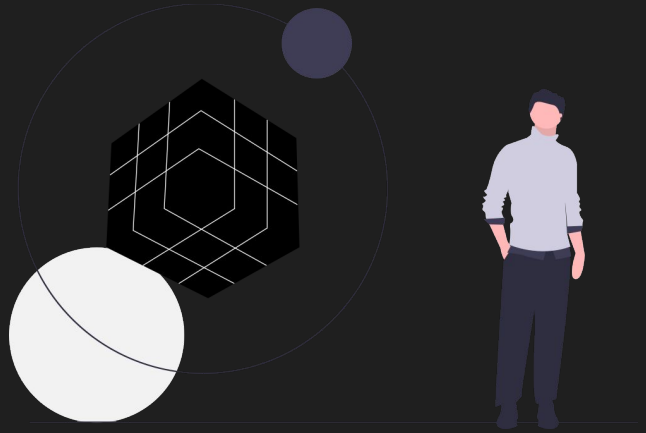
2022

SEAZONE CHALLENGE

WHAT SEAZONE DATA TELLS ABOUT ITS OWN FUTURE?



SUMMARY

1. QUESTION 1
2. QUESTION 2
3. QUESTION 3
4. QUESTION 4
5. PROBLEM - SOLUTION CYCLE
6. CONSIDERATIONS

QUESTION 1

LISTING PRICING

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

QUESTION 1

LISTING PRICING

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

The expected value for a random variable is better estimated by its average over observed values. However, there are no listings in the provided dataset presenting all 3 required characteristics.

Therefore it is necessary to train a regression model for predicting price and revenue given location, category and number of rooms. Two models were trained: a linear regressor and a XGboost.

The XGBoost model has a lesser out-of-sample RMSE, nevertheless it predicts **R$ 312,20** for the listing: a not so great prediction since it is known from real data that JUR MASTER 1Q and JUR MASTER 3Q average prices are, respectively, **R$ 383,00** and **R$ 531,64**.

Given that the linear regressor predicts **R$ 586,21**, an ensemble model is made out of both of them. Testing different weights yields a 80/20 best balance in favor of XGBoost. The ensemble predicts a **R$ 367,00** price.

Even though it is still below the JUR MASTER 1Q, this prediction is favored for considering more information in it. Seazone revenue is, therefore, estimated as the owner revenue multiplied by the commission fee. The commission fee is adopted as 20% since JUR MASTER 1Q, 3Q and 4Q all share this same value for it.

Final answer

Owner price and revenue: **R$ 367,00**.
Seazone revenue (commission): **R$ 73,40**.

OWNER REVENUE COMPARISON

Month	Location	Category	Bedrooms	Average Revenue
mar	JUR	MASTER	1	R$ 383,00
mar	JUR	MASTER	3	R$ 532,00
mar	JUR	MASTER	4	R$ 2.224,00

QUESTION 2

REVENUE FORECAST

What is Seazone expected revenue for 2022? Why?

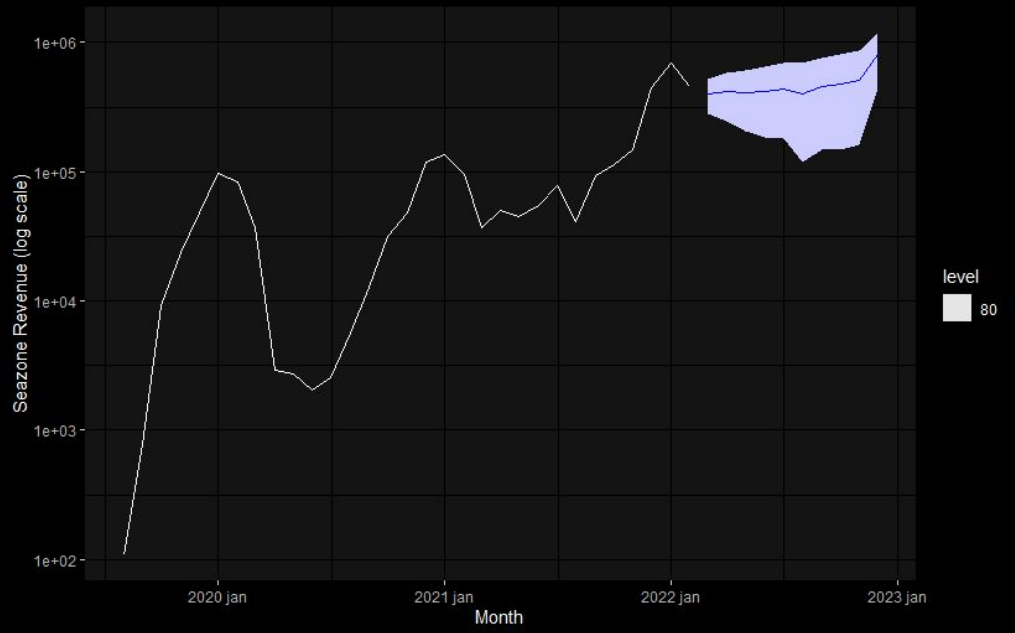
QUESTION 2

REVENUE FORECAST

What is Seazone expected revenue for 2022? Why?

Through accuracy assessment, a Seasonal Autoregressive Integrated Moving Average model – SARIMA(1, 0, 0)(0, 1, 0)[12] – was chosen to fit the monthly revenue data derived from Seazone total commissions earned up until 03-15-2022.

The forecast suggests a total revenue in 2022 in an 80% confidence interval ranging from R$ 3.261.318,82 to R$ 8.484.595,04 with expected value **R$ 5.872.956,93**.

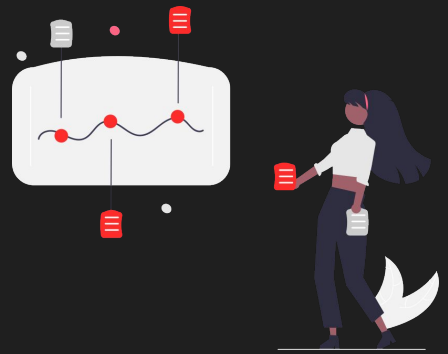


CLICK THE PLOTS TO LOAD AN INTERACTIVE VERSION

QUESTION 3

DEMAND FORECAST

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



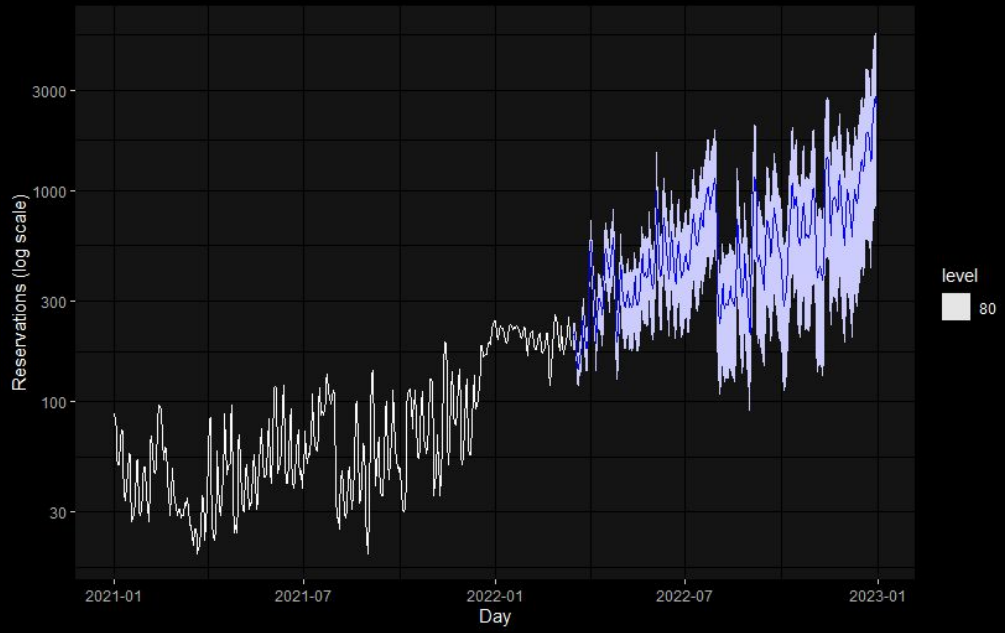
QUESTION 3

DEMAND FORECAST

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

To forecast daily data with complex behaviour, a model that can capture multiple seasonalities and is robust to sudden changes (like the pandemic) is necessary. Also, as the seasonality effects scales with the reservations made each day, a log transform should be applied prior to modeling. For these reasons, a dynamic harmonic regression model is fitted using Fourier terms to account for seasonality, an ARMA model to account for short-term dynamics and a dummy variable "is workday?" as an exogenous predictor.

The **results for each day** may be seen by clicking the **plot to the right**.



CLICK THE PLOTS TO LOAD AN INTERACTIVE VERSION

QUESTION 4

NEW YEAR'S NIGHTS

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

QUESTION 4

NEW YEAR'S NIGHTS

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

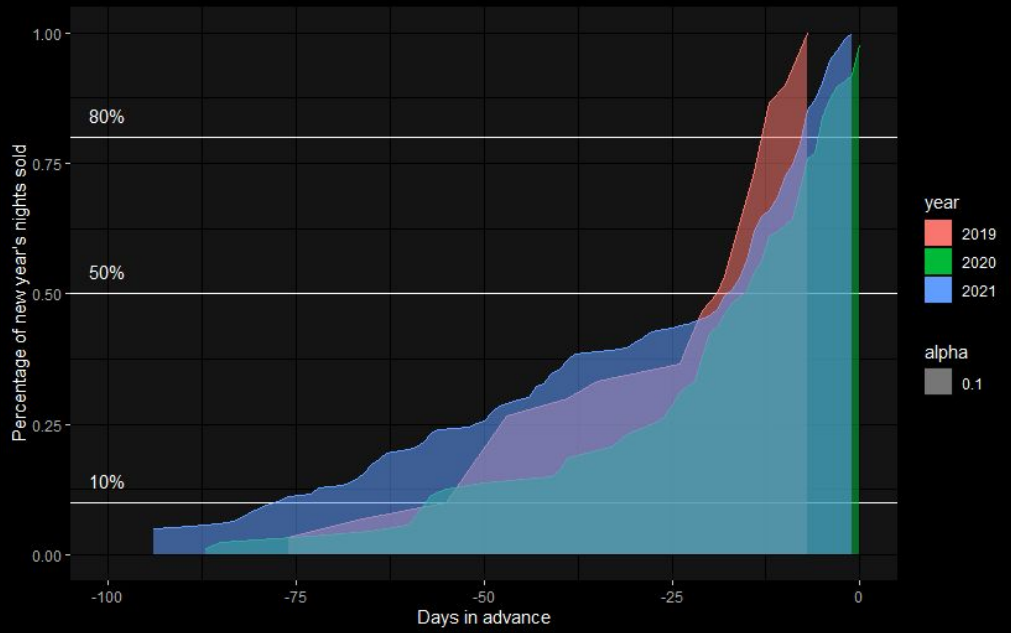
Averaging over the days in advance in which each of these marks were reached in 2019, 2020 and 2021, results in:

29/10, 14/12 and 23/12, respectively.

Pricing is considerably dependent on reservation advance since it carries a trade-off: on one hand, a customer is willing to pay more for booking to a close date; on the other, there is a strong incentive to sell as many bookings as possible up to the target date.

To set an optimal price is to reach an agreement between what is mutually interesting for both the business and its clients.

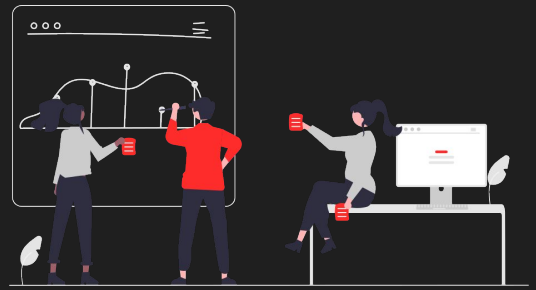
Furthermore, a model for the odds of selling a booking given the percentage of available listings for a particular date at a particular price may be used to maximize revenue as a function of price.



CLICK THE PLOTS TO LOAD AN INTERACTIVE VERSION

SUMMARY

- QUESTION 4
- **PROBLEM - SOLUTION CYCLE**
- CONSIDERATIONS



ANALYSIS MEANS

STACK CHOICE

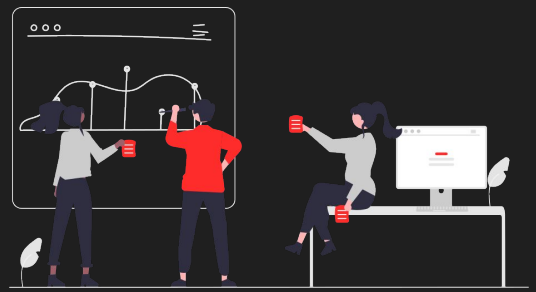
Initially, R was considered only for the EDA and data cleansing stages of the analysis due to the excellent integration between dplyr and ggplot packages and their versatility. Thus, Python would be applied in the modeling stages making use, mainly, of numpy, seaborn, matplotlib and scikit-learn libraries.

However, it became clear as more time of the analysis was spent in the time series forecasting stages that a better approach was to make use of tidyverts, a great family of R packages to manipulate TS data.

Whichever of those ways were taken, Chart Studio would be used to report results due to its easy integration to Python and/or R scripts.

SUMMARY

- QUESTION 4
- **PROBLEM - SOLUTION CYCLE**
- CONSIDERATIONS



ANALYSIS MEANS

EDA

All EDA code and steps are documented in the R Markdown file ExploratoryDataAnalysis.Rmd located in the scripts folder from the repository.

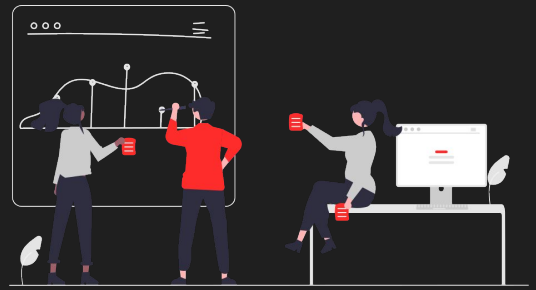
A key regard in this stage was to develop a great understanding of the data at hand. To this goal, two models were fitted and the predictive power of each selected variable was assessed.

Missing values were a problem for three main reasons:

1. Hotel rooms have slightly, but sensible, different characteristics compared to apartment and house listings (like having no number of bedrooms), so a split hotel/no-hotel was made in the data;
2. Number of pillows was missing more than other variables, but it was discarded for the model fitting stages since it is highly correlated with capacity;
3. A specific listing “TSTOO1” appeared in the daily revenue data, but it was note present in the listings database. It was discarded since no revenue was ever made from it.

SUMMARY

- QUESTION 4
- **PROBLEM - SOLUTION CYCLE**
- CONSIDERATIONS



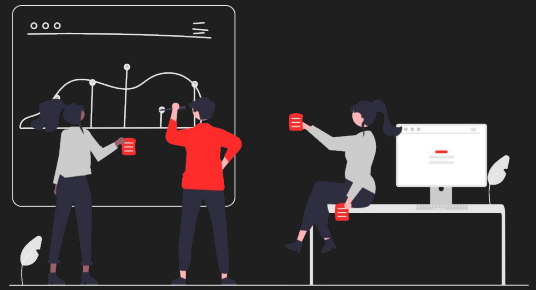
ANALYSIS MEANS

EDA

Finally, cleaning fees, reservation advance and dummies for holidays, location and categories ranked first as the most relevant price and revenue predictors.

SUMMARY

- QUESTION 4
- **PROBLEM - SOLUTION CYCLE**
- CONSIDERATIONS



ANALYSIS MEANS

REVENUE MODELING

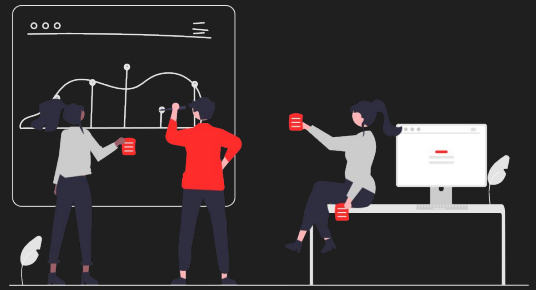
All Revenue Modeling code and steps are documented in the R Markdown file Modeling_Revenue.Rmd located in the scripts folder from the repository.

A initial decision to be made was to fit a model on daily or monthly data. As the monthly aggregation smooths out some inconsistencies, it was preferred.

ETS and ARIMA models were tested using both manual and automatic procedures for parameter selection and with or without a prior logarithmic transform. The best fit found through cross validation was a SARIMA(1, 0, 0)(0, 1, 0)[12] model.

SUMMARY

- QUESTION 4
- **PROBLEM - SOLUTION CYCLE**
- CONSIDERATIONS



ANALYSIS MEANS

DEMAND MODELING

All Demand Modeling code and steps are documented in the R Markdown file Demand_Revenue.Rmd located in the scripts folder from the repository.

Right away, due to the complex seasonalities, the chosen model was a dynamic harmonic regression. Some experimentation was made to choose which exogenous variables to use and eventually, mainly due to computational troubles, a single dummy variable (is.workday) was selected.

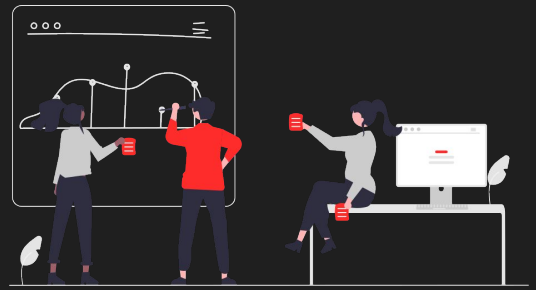
Most time here was spent in the K hyperparameter tuning. K is the number of pairs of sine and cosine functions in the fourier harmonics expansion and, interestingly, presented more than two local *minima* for the time series cross-validation performance metrics.

The need for a pre-applied logarithmic transformation was tested here as well.

A prophet model could be chosen as well, but it was discarded since it is known to make worse predictions in most cases.

SUMMARY

- QUESTION 4
- **PROBLEM - SOLUTION CYCLE**
- CONSIDERATIONS



ANALYSIS MEANS

CHALLENGE ANSWERS

All Challenge Answers code and steps are documented in the R Markdown file ChallengeAnswers.Rmd located in the scripts folder from the repository.

Armed with a good understanding of the data and good models, the process of answering the questions felt natural yet challenging and rewarding.

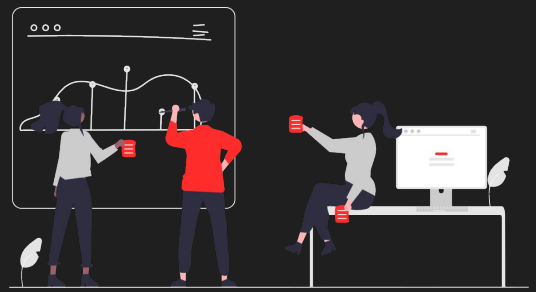
Special attention was devoted to selecting a good predictor for Question 1 as already mentioned.

A simple model for answering Question 4 was favored due to the high odds that a complex one would produce very similar and maybe worse results.

Question 2 and 3 were already answered by the time the modeling stages were concluded.

SUMMARY

- QUESTION 4
- PROBLEM - SOLUTION CYCLE
- CONSIDERATIONS



FINAL CONSIDERATIONS

NEXT STEPS

- Enhance models developed in this challenge using techniques from the paper "Prediction accuracy for reservation-based forecasting methods applied in Revenue Management" by Fiori and Foroni that suggests a novel stochastic pickup method that uses already known booking data to predict demand and revenue.
- To test a daily approach to revenue prediction and compare to the monthly approach.
- To test a prophet model against the dynamic harmonic regression.
- Use a ridge/lasso regression in the EDA stage of identifying relevant predictors for revenue

CHALLENGE FEEDBACK

Working with real-life data to tackle business problems is always a blessing. I highly appreciate the effort to state the questions clearly and explain well what each variable meant.

My best regards to all the Seazone team.

"Without big data, you are blind and deaf and in the middle of a freeway"

– Geoffrey Moore

WE WANT YOU TO GET THE MOST OUT OF EACH DESTINATION AND LIVE, INTENSELY, NEW STORIES ON THIS BIG WORLD.

ANALYSIS

BY GUSTAVO MELLO