Lighthouse technical exercise

Senior Data Scientist - José Rodrigues!

Before we meet again for the technical interview, we would like to ask you to show us how you deal with the data when you're at ease. The best way to do this is - what a surprise - a little exercise to prepare at home. You will be solving a real problem we dealt with at Lighthouse, using a real-life dataset!

Good luck :)

Guidelines

It is not a start-to-end project that you need to solve - it's just an exercise to show us how you approach the problem and work with the data. So please don't spend too much time making it perfect - a few hours should be enough. This is also not an exam, you should reach out to us if something is not clear or if you are stuck.

Use your preferred technology to go over the tasks - could be Python, R, SQL, Tableau, Excel, Paint or whatever as long as you provide us your “code” and your results.

In the "Tasks" part of the doc we described the exercises. Please use this part also to describe your findings - try to describe your thought process and write the conclusions as clear as you can. All charts to visualise the results are more than welcomed.

Additionally, put the link to your code at the end of this doc or send it zipped by mail, so we can see what tools you're used to working with and how your code looks. You really don't have to worry about the clear code rules!

Data model

We've attached two files, and here are short descriptions of the schemas:

hotels_information.csv
This file is a set of basic facts about hotels. One row describes one hotel.

Field	**Type**	**Meaning**
our_hotel_id	integer	Unique identifier of the hotel.
name	string	Name of the hotel.
review_score	float	Review score of the hotel.
stars	integer	Star rating of the hotel.
latitude	float	Latitude of the hotel's location.
longitude	float	Longitude of the hotel's solcation.
room_count	integer	Number of rooms in the hotel.

pricing_data.csv
In this file we included pricing data for the set of hotels from the first file. We included an extract of the prices for the next 90 days, generated on 2021-01-22. For simplicity, for each hotel we selected one representative room. There is no information about the currency, but you can assume that it's € everywhere.

Field	**Type**	**Meaning**
our_hotel_id	integer	Unique identifier of the hotel.
arrival_date	date	The date of stay.
lead_time	integer	The time (in days) between the sample generation date and the stay date.
room_name	string	Name of the selected room.
meal_type_included	string	Type of the meal included in the offer.
max_persons	integer	Number of persons allowed in the room.
price_value_ref	float	Price value for the refundable offer.
price_value_non_ref	float	Price value of the non-refundable offer.
is_sold_out	bool	True/False if the hotel is fully sold out.

Tasks

1. Data discovery

Get familiar with the datasets. Describe what kind of things you looked at and why. Document your findings.

2. Markets comparison

2.1. The dataset contains hotels in two different markets. Figure out what they are and split hotels into two groups.

2.2. Using all the available data, **compare** the two groups in the following aspects:

2.2.1. How many hotels are available **per arrival day** in the next 90 days?

2.2.2. What are the overall pricing patterns?

2.2.3. What are the distributions of hotels when it comes to stars and review scores?

2.2.4. [optional] Anything else you could think of that would be worth exploring?