# **CASE STUDY** – Data Scientist ( Source)

**INTRODUCTION:**

Our mission at trivago is to provide travelers with a high-quality accommodation search tool. To aid in this mission, the Source domain processes a wide variety of data, including reviews, descriptions, images, and locations. We enrich and restructure these data to provide our customers with the most relevant possible information on our website and apps.

One of the ways we add value is to tag accommodations with experience-related concepts. In this case study, we would like you to tackle one such concept (nightlife) for some hotels in a few selected cities. Show us your creativity and hands-on data science skills by creating an algorithm/model that ranks the given hotels for the given concept.

**SUBMISSION DEADLINE:** 7 days from receipt

**HOW TO SUBMIT:** Please submit your document via the link sent by your recruiter in the email. The format is your choice!

## THE CHALLENGE

### Task 1: Exploratory analysis of the input data ( ~15% time)

In the resources section you will find two CSVs containing the following data (the detailed data schema is given in a PDF):

- *hotels.csv:* Basic information about hotels from four different cities (Amsterdam, Los Angeles, Hong Kong, Thessaloniki), including their geo-coordinates, some information about the type of the hotel and the level of exposure they received in the past.

- *pois.csv:* Basic information about POIs (points-of-interest) in the same four cities, including the POI's geo-coordinates and the types of the POIs.

**Perform exploratory data analysis** on these datasets using a language of your choice. Get a feeling for what information is contained in it. Since you will be working with geographical data, you could for instance show off your visualisation skills by plotting some interesting maps with this dataset. While doing this analysis, think about what information in this dataset could be useful for assessing a hotel's suitability for travelers who are mostly interested in the nightlife aspect of the city.

The expected output is the code you used to do your analysis and a report with visualizations of the data. The format in which you deliver them is up to you (e.g. one Jupyter notebook including your code and report, a text i.e. with your code and a PDF with your report etc.).

**Task 2: Creating nightlife scores for hotels ( ~55% time)**

Using the data provided come up with **an algorithm or model that quantifies the suitability of a hotel for nightlife travelers in each city.** In other words, we would like you to assign a nightlife concept score to each hotel. For instance, you might create scores between 1 and 100, where 100 is the best possible nightlife score. Be sure to give some explanation as to the interpretation of the score, i.e., what's the difference between a hotel with a nightlife score of 20 versus 80 versus 90?

For this task you will have to make some assumptions about what data should influence such a suitability score (i.e., what will be the features). It is absolutely fine to make these assumptions as long as you document them in your deliverable and add some explanation so that we can follow your way of thinking. For example, you could assume that the proximity to a pub or bar influences the suitability positively, i.e. the closer a hotel is to a bar, the better its nightlife score should be. This is intentionally a very open task (which features to use, how to combine them, what scoring algorithm etc.) and there are many possible approaches that can produce the desired output. Make some choices even if you are uncertain about them, but do document your reasoning as clearly as possible.

The expected output is the code you used to create the scoring algorithm/model and some documentation/explanation of your reasoning. In addition, the scores you have created should be delivered in a csv with the following columns: *hotel_id, city_id, score*

**Task 3: Presentation of the nightlife scores (~35% time)**

Now imagine that you have produced nightlife scores for all hotels on trivago using the algorithm you created in Task 2. A product manager approaches you because they would like to build a feature that uses the scores to rank hotels when a trivago user explicitly states their interest in the nightlife of a destination. You would like to convince them that your scores are ready for this feature. We would like you to prepare a presentation to support your point.

Make sure you mention the following in your presentation:

- Overview of the algorithm/model you created
- Explanation of how you validated the scores
- Plans for how you would implement and test the feature
- KPIs for the feature once it is rolled out
- Maintenance plan for the feature

**Target a presentation time of 20 minutes. You will be delivering this presentation in the next round of the interview if your case study is convincing, so also be prepared for questions regarding it.**

***General remarks for the case study:***

This case study is formulated in a very open manner, and there is the potential for spending a lot of time on a submission. Make sure ththat you timebox yourself and don't go overboard with experimenting with the data. If you're short on time, focus more on Tasks 2 and 3