# CASE STUDY –

## Data Scientist in Data Products

Hi and welcome to the trivago Data Scientist - Data Products recruiting data challenge.
Your application made us curious and therefore we would like to take you to the next step.
This is the chance for you to have a glimpse into our data and to convince us that you are the right person for the job.

We wish you good luck!

**Estimated work time:** 8 hours

**SUBMISSION DEADLINE:** 7 days upon reciving the case study.

### THE CHALLENGE

In trivago, we keep a log of all attributes and user activities within a session and keep them in tabular format with each row representing one session.
Due to some technical problems, the data was corrupted for the period of a week and information about “hits” (the number of interaction with trivago page) is lost for some sessions.

Along with this document we have provided you with data in a .csv format containing information of about six hundred thousand sessions from this period.

1. Using the data provided, please answer the following questions:
    a) Are all locales impacted by the loss of data equally?
    b) Which 5 path_id have the highest average hits in each locale? How about globally?
2. Note that the column “hits” has missing values. Use this data to build a model that predicts the number of hits per session, depending on the given parameters.
    a) What other metrics can your model predict that can be useful?
    b) What other columns would you like to have to improve your model?
    c) Can your model predict the hits for tomorrow?

**Description of the dataset:**
The columns should be understood as follows:

- index: a number uniquely identifying each row.
- locale: the platform of the session.
- day_of_week: Monday-Friday, the day of the week of the session.
- hour_of_day: 00-23, the hour of the day of the session.

- agent_id: the device used for the session.
- entry_page: the landing page of the session.
- path_id_set: a set of all the locations that were visited during the session.
- traffic_type: the channel the user came through eg. search engine marketing, email, …
- session_duration: the duration of the session in seconds.
- hits: the number of interactions with the trivago page during the session.

**Factors that we use for evaluating the results:**

- Your ability to extract meaningful insight from the data and your answers to the questions.
- The code quality, the steps you went through and the techniques you used to create your model.
- Your predictions will be evaluated by the root mean square error:

$$error = \sqrt{\frac{\sum_{i=1}^{n}(Predicted\ Hits_i - Observed\ Hits_i)^2}{n}}$$

**How to Submit:**

1. A pdf or powerpoint containing answers to the questions together with your thought process or any interesting observations.
2. The code/notebook you used for this task.
3. For your predictions, a .csv file containing two columns: index and hits for all the rows where hit data is not available.