# Coding Challenge for Data Scientist with Statistics focus Candidates at Babbel

## General remarks

Data Scientists at Babbel are not developers, and a lot of heavy lifting is done by our great data engineering team. Nonetheless analysts must be able to code at a more than basic level to be effective. Analytics at Babbel is not an Excel job!

This coding challenge is based on actual data and business problems from Babbel. You may solve it either using the **Python** data stack or using **R**.

At Babbel, analytics – as well as data science and data engineering – is based on Python (Python 3, Jupyter, Pandas, PyMC, scikit learn, etc.). If you are coming from an R background, you can start the position using R for your analyses, but we expect you to learn Python in the long run.

If you solve the challenge using **Python**:

- Please submit a Jupyter notebook. Feel free to use graphics and text to explain what you did.
- You may use any package you like, as long as it is installable via `pip` or `conda`.
- Please use Python 3.

If you solve the challenge using **R**:

- You may use any package you like, as long as it is installable from CRAN.

Some general remarks:

- We like clean, non-redundant code that makes use of the features of the programming language you chose, without going crazy.
- Comments are good!
- So are plots!

# Data Analysis Challenge

- Please use the data file `learner_item_data.csv` that comes with the challenge documentation.

The dataset `learner_item_data` describes events that are recorded when a user does vocabulary training.[1] Each event corresponds with a user solving one “trainer item”, which is a vocabulary question, a grammar task or something similar. During a training session, a user will solve a bunch of trainer items. This is the structure of the data:

| **column name** | **semantics** |
|---|---|
| `uuid` | user id |
| `created_at` | timestamp of event creation |
| `trainer_item_it` | ID of the review question |

Please provide the code for the following analyses.

1. The events are not grouped into sessions, so your first task is to identify those. We define a session as a sequence of events with the same user ID, ordered by timestamp, such that the time difference between any consecutive pair of events is at most one hour.
   Your task is to identify the sessions of each user and encode them in the dataset in a way that will allow you to evaluate statistics on sessions.

Example:

| **`uuid`** | **`created_at` (just showing the time)** | |
|---|---|---|
| 42 | 16:50 | one session |
| 42 | 16:55 | |
| 42 | 17:02 | |
| 42 | 20:32 | another session |
| 42 | 20:35 | |

[1] The dataset is real, but the IDs are not the ones actually used in the backend.

2. Using the sessions, answer the following questions. Use plots and text as you like to communicate your results.
    - What is the distribution (histogram) of session lengths?
    - How many people do several sessions a day? How many sessions?
3. Assume Babbel built a new feature that is supposed to increase activity. To test if it actually increases activity it got released as an A/B test with 40% of the users seeing the new feature (test group) and 60% not seeing it (control group). You agreed with the Product Manager on two metrics to measure the activity:
    - Proportion of users per experiment group who do more than one session on at least one day
    - Median time per session

    You can find the assignment of users to experiment groups in the file `test_groups.csv`. Using the activity data in `learner_item_data.csv` and the assignment answer the following questions:
    - Is there a difference between the experiment groups in any of the two activity metrics?
    - If there is a difference, how big is it?
    - What would you recommend the Product Manager to do next?