This is an ABSA open-data task for Data Science position candidates, please solve this task on your own and do not share the task or the solution with anybody. Please send us your solution within a week after receiving this task.

We have provided you with 2 CSV files and a set of questions. This task is designed to evaluate the level of your knowledge and technical skill. The questions are intentionally ordered from the simplest to ones suited for more experienced candidates. Please try to solve as many questions as you are able to, but don't be concerned if you are not able to progress past a certain point.

You can solve the given task using the Python programming language in any environment that suits you. If you're not comfortable with Python you can use the language of your choice but note that we would like to see some plots and charts. (Do not attempt to solve this task manually or in Excel-like applications!)

Please make your solution clear, with comments and notes so we can understand your process. If you feel some questions are vague and have multiple directions, please make a decision and defend them in your commentary.

Should you pass our review, you will be required to present your solution in an interview and maybe be questioned on your approach and decisions, so please remember to take notes and prepare your solution for the interview session if invited.

Please do not spend an excessive amount of time on this task, that is not the point. We would also like you to keep a rough estimate of the time spent and ask you to provide it to enable us to improve our interview processes.

Best of luck!

Dataset

Our dataset is composed of 2 files: a candidate information file and an admissions file for an imaginary company.

Candidate file

- name
- candidate_id
- age
- gender
- no_children - number of children
- emp_card_id - id of employee card

Admission file

- candidate_id
- department
- admit

Questions

Our set of questions is intended to build on itself with an ultimate goal of training a model that would predict if a candidate will be hired.

Keep in mind that we as ABSA data scientists will often need to explain our choices to both our colleagues and people without education in the field.

1. Determine an average age of a candidate
2. How is the data quality? How do you check for it?
3. Join both datasets.
4. Perform a feature selection process on given data for the model in the next question.
5. Train a model that given selected features predicts the likelihood of a candidate being admitted.
6. Verify the model you have trained using a metric of your choice.
7. Is there any discrimination going on in the company's hiring?

- What is the rough estimate of the time you spent on this task?