# IA|BE Data Science Certificate - Module 1

## Edition October 2021

### Instructions for the Assignment

You should provide answers in Dutch, French or English. You should demonstrate in your solution that you master the methods that have been discussed in Module 1. You are not supposed to use any approaches or methods that have not been covered in the Module 1 sessions. You must use `Python` for your calculations and graphics, in line with the computer labs of Module 1. Occasionally, and if justified, you can call a specific `R` library from `Python`.

Success!

### Deliverables for the Assignment

Please hand in on or before January 16, 2022 via email:

1. A notebook (.ipynb) or link to a Google Colab that documents your modelling steps and provides guidance for the reader, including some discussion of your findings and obtained insights. Your code should be well-organized and easy to read.

Please mention the names of your team members on both items. It is allowed to work in teams (with two students maximum); it suffices to submit one solution per team.

Each team of students will deliver an (online) pitch presentation (schedule to be determined). This allows the teaching team to give feedback on the report and the models constructed.

### Assignment Questions

You analyze the data set (in `.csv`) that is available on the home page of Module 1. This data set contains observations on the variables listed in the table printed below. Your report should document the following steps:

1. An exploratory data analysis.
2. The construction of a (technical) tariff structure for a car insurance product. Hereto you analyze either the frequency or the severity information in the data set with (at least) two of the methods/algorithms discussed in the Module 1 lectures (e.g. GLM, GAM, neural network, regularized GLM). Alternatively, you can build a model for frequency and a model for severity (e.g. a GLM for frequency and a GLM for severity, or a neural net for claim frequency and a GLM for severity). You discuss the essential insights obtained with these models. You compare the performance of the constructed models, based on your own defined set of criteria. You define some risk profiles and illustrate the predictions obtained with your models for these risk profiles.

There is no need to answer the above questions separately (question by question) in your report. A well structured text that covers the above items is preferred. Be creative and rigorous! While the data set lists the 4–digit postal code of the policyholder, we do not expect a detailed analysis of the spatial heterogeneity (as this goes beyond the methods explained in Module 1).

| | |
|---|---|
| `ageph` | age of the policyholder |
| `CODPOSS` | postal code in Belgium |
| `duree` | exposure, fraction of the year the insured is covered |
| `lnexpo` | log of exposure |
| `nbrtotc` | total number of claims during period of exposure |
| `chargtot` | total claim amount |
| `agecar` | age of the car: $0-1$, $2-5$, $6-10$, $>10$ |
| `sexp` | sex of the policyholder: male or female |
| `fuelc` | type of fuel: petrol or gasoil |
| `split` | split of the premium: monthly, once, twice, three times per year |
| `usec` | use of the car: private or professional |
| `fleetc` | car belonging to a fleet: yes or no |
| `sportc` | sport car: yes or no |
| `coverp` | coverage: MTPL, MTPL+, MTPL+++ |
| `powerc` | power of the car: $<66$, 66-110, $>110$ |

Katrien Antonio, Jonas Crevecoeur, Roel Henckaerts, Michael Lecuivre and Samuel Mahy.
Version November 2021.