# Lecture 04: Building computer models with visual analytics Exercise description

## Goals

Use visualisations in creation, evaluation, and refinement of computer models. This is done by example of univariate regression models, which is a very simple model type. The focus of the exercise is not the modelling by itself but application of the principles of conscious model building, i.e., involvement of analytical reasoning throughout the process.

## Data

The exercise is done using census data for the London wards. The file population_perc.csv contains attributes describing the population structure as percentages of different population groups based on the age, sex, nationality, occupation, qualification, and other parameters. There are a few other attributes, such as the mean age and the average distance to work.

You are expected to explore and model the interrelations between the qualification level and the health condition. For each of these aspects, there are several attributes specifying the percentages of population in different categories by the qualification level and health condition. We propose you to look at the extreme categories: no qualification or high qualification versus very good or very bad health. The respective attribute names are:

- Qualification:
  - "*qualification (study)=**No qualifications**: Population % by qualification or study*"
  - "*qualification (study)=**Level 4 qualifications and above**: Population % by qualification or study*"
- Health condition:
  - "*health condition=**Very good health**: Population % by health condition*"
  - "*health condition=**Very bad health**: Population % by health condition*"

## Draft Python scripts

We provide you with two draft scripts in Python (as Jupiter notebooks). You use these scripts as a starting point, but you are expected to adapt them to the tasks at hand along the analysis process.

Script contents:

- 04-1-initialExploration.ipynb: Data loading, joining the population data with the geographic boundaries of the wards, removing wards with missing values, checking the link between the geographic data and the population attributes.
- 04-2-model.ipynb: Fitting a regression model, variation of the model parameter (order), looking at model residuals, and construction of multiple models for different divisions of the input data.

## Tasks

- Explore visually the interrelations between the attributes referring to the qualification and health condition; choose an attribute pair for building a formal model (regression) describing their interrelationship.
- Build model variants for different values of the parameter (order) and assess visually which variant gives a better fit.
- Consider the distributions of the residuals for the model variants. Choose the variant with the best (i.e., the most random) distribution of the residuals. If two or more models are equivalent in this

respect, choose the simpler one. How does your choice correspond to the previous visual assessment of the goodness of the model fit?

- Check whether the interrelation varies over the territory.
    - Build models for the data subsets based on the attribute 'Borough' and look for existence of significant differences.
- Check the possibilities to refine the model using some attributes as conditioning variables.
    - Try attributes "Mean age" and "Average distance to work". Which attribute has a better potential for refinement (i.e., the partial models differ more significantly)?
- Check the possibility to refine the model by using clustering based on several attributes reflecting the age structure, in particular, proportions of the age groups from 25 to 64 years old, which seem the most relevant to the qualification level.
    - Apply partition-based clustering and build a model for each cluster. Are there essential differences?
- Based on the analysis done, choose a reasonable approach to modelling: to build a single model or several partial models; if several, how to divide the data; what order to use. Take into account model accuracy, complexity, and understandability.

We suggest you to note your findings as comments in the notebooks you are using and to share your notebooks with the changes and notes you have made in the moodle forum.

We wish you a successful and fruitful fulfilment of the exercise.