# Project 2

Barbara Holland

# Dataset

Refer to the dataset ('Bank.csv') that can be found on MyLO. The file contains data on 93 employees of a bank who all belong to the same job category (skilled, entry-level clerical). The employees (32 male, 61 female) were all hired in the period 1965-75. For each employee for following things have been recorded.

- `Bsal` - Base salary at time of hire;
- `Sex` - M (males), F (females);
- `Senior` - Months since first hired;
- `Age` - Age in months;
- `Educ` - Years of education;
- `Exper` - work experience prior to job with the bank (months).

# Question

The bank was sued for sex-discrimination. To prove discrimination, it's not enough to just notice that there is a difference in average salary between male and female employees. We are interested to see if there is evidence of discrimination after potential confounding variables such as experience and seniority have been taken into account. In particular: Did females receive lower starting salaries than similarly qualified and experienced males?

# Analysis tasks

1. Visualise the relationships between `BSal` and the potential non-discriminatory explanatory variables `Senior`, `Age` `Educ` and `Exper`. Does this reveal any interesting features of the data or give you any ideas about which variables may/may not be important to include in the model? Does it give you any ideas about whether or not transforming variables may be helpful?
2. Construct a preliminary model for the non-discriminatory variables and use it to assess if any transformations of the data may be required and if any outlying points need to be dealt with.
3. Work through a strategy for variable selection that addresses the question of whether there is any evidence of discrimination with regard to starting salary.
4. Assess how well your final model fits the data - do you have any remaining concerns about the diagnostics?
5. Given your model, what are your conclusions regarding any link between `Sex` and `BSal`? How robust are your conclusions to different choices about model selection?

# Writing the report

Write a report with two sections. The first section should:

- Summarise your conclusions for a general readership, and
- Include a graphical representation of the data.

The second section should:

- Detail your statistical methodology,

- Provide the relevant output of different models (with commentary),
- Include diagnostic plots (with commentary)
- State why and how you reached your final conclusions.

Have fun! Ask lots of questions if anything is unclear :)

**The marking rubric for the project can be found on MyLO. (It is the same as for Project 1)**