Dataset

Find the dataset `ex1029` that is part of the `Sleuth3` package.
(run `library(Sleuth3)` followed by `head(ex1029)` to have a look) The file contains information on the weekly wages of 25,437 American males from 1987. For each person the following things have been recorded.

- `WeeklyEarnings` - Weekly earnings in $;
- `Region` - Midwest, Northeast, South, or West;
- `MetropolitanStatus` - Whether or not they live in a metropolitan area;
- `Exper` - Years of experience;
- `Educ` - Years of education;
- `Race` - coded as a binary categorical variable (Black or not).

Question

The key question of interest is whether or not race affects weekly earnings. To show discrimination, it's not enough to just notice that there is a difference in average wages between black and non-black workers as there may be many confounding variables. In this case several potential confounding variables have been recorded. We are interested to see if there is evidence of a difference in wages between blacks and non-blacks after these confounding variables have been accounted for. In particular: Did black men receive lower wages than similarly educated and experienced non-black males?

Analysis tasks

1. Visualise the relationships between `WeeklyEarnings` and the potential non-discriminatory explanatory variables `Region`, `MetropolitanStatus` `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 racial discrimination with regard to wages.
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 `Race` and `WeeklyEarnings`? How robust are your conclusions to different choices about model selection?