CI7330 – Data analytics and visualisation

Autumn term 2021: formative assessment

Your task is to carry out simple data analysis of a dataset and submit a short report, where *the audience is intended to be people with no experience of statistics or data analytics*.

The dataset you have been given is taken from the National Labor Survey for Women in the USA. In particular, we are looking at the year 1988. The dataset has been altered from the original, so you will not find any "answers" by reading other people's analyses of it online.

There are the following variables:

- id: a number identifying each woman in the dataset
- age: a quantitative variable taking integer values between 34 and 46 (because of the way the survey followed cohorts of women over time)
- race: a nominal variable with the following values:
    - 1 = white
    - 2 = black
    - 3 = other
    - (this reflects the simplified way that racial data were used in 1980s America; now, we would want to know more information, and not just to put people under "Other")
- collgrad: a binary variable, 1 if the woman graduated from college / university, 0 otherwise
- south: a binary variable, 1 if the woman lived in the Southern states, 0 if in the North
- ttl_exp: a quantitative variable taking any positive real value: the total years of experience in the workplace
- hours: a quantitative variable taking positive integer values: how many hours per week the women had worked in the previous year
- ln_wage: a quantitative variable taking any positive real value: the logarithm of the wage per hour (the logarithm transformation helps to make the variable symmetrically distributed).

You need to submit:

- A Word document clearly marked with your student number but not your name.
- This document should contain three parts, headed "Part 1", "Part 2" and Part 3"
- Part 1 should contain one table of appropriate descriptive statistics that summarise the variables. Note that some variables are categorical and some quantitative. If you submit more than one table, only the first one will be marked.
- Part 2 should contain two visualisations that show aspects of the data that you think are interesting. You are free to choose what these are. One must show a *single* variable, and the other must show the relationship between *two* variables. There should be a short caption under each of the images. If you submit more than two visualisations, only the first two will be marked. You can use any software for this.
- Part 3 should be a text description of the analysis and what it shows. It should be written in an informal style with no Aims, Methods or other scientific headings. Remember that the audience are not experienced in statistics or data analytics. This

Part 3 must not be any longer than 500 words. It must explain what the reader can see in Parts 1 and 2.

Deadline:
This assessment must be submitted by Friday 3 December at 12:00.

This is a formative assessment: you will receive feedback on it, but it will not count towards your final mark for this module. However, you must submit it to be recorded as having taken part in the module constructively.