SCIF30006 – Advanced Data Science for Scientific Computing

Week 2, Workshop 1 – Exploratory Data Analysis

Please bring your laptop or another device suitable for coding with you to this workshop. You'll need the Python/Jupyter IDE of your choice and (optionally) RStudio.

Get together in groups of 3. Ideally, arrange for groups to include people with different preferences for data analysis packages, including somebody willing to have a go with RStudio.

Download the datasets for this workshop from BB. You should find an Excel file containing 2 sets of exam marks and also the individual sheets as csv files.

Start out by inspecting both data sheets – personally, I would probably use Excel to do this, because the datasets are small, but it is also a good opportunity to remind yourselves of other commands.

Q1 Do you see any problems with this data as recorded?

Read data1.csv into the package of your choice.

Generate the following descriptive statistics:

- Sample size.
- Minimum and maximum values, range for the individual questions and the total marks.
- Median and mean marks.

Use these data to answer Q2-Q4:

Q2 Compare the median and mean marks. Can you see much difference between these two measures?

Q3 Was the full range of marks accessible in this case?

Q4 If you had set this exam question, would you be happy with the outcome?

Now read data2.csv and perform the same analyses. Compare the descriptive statistics for both sets and decide whether you want to flag any scripts for review.

Answer Q2-Q4 for the second set of exam results.

Try out different plots to decide how best to present these exam results to an exam board meeting. Compile a summary of key results using no more than one side of A4 for each exam paper. Compare your summary with that of other groups.