CASA0006: Data Science for Spatial Systems
Assessment Guidelines

Deadline 5pm, 21 April 2025, Monday, UK Time
Word Count Maximum 1500 words (not including Python scripts or comments)

The coursework for this module will consist of an individual assignment that tests your ability to conduct in-depth data analysis. Each student is required to submit a single Python Notebook which contains both the code required to conduct the data analysis and accompanying text which provides context interpretation.

This coursework represents 100% of the overall module assessment.

Task

Please choose one of the following datasets as the main datasets and define a research question, which should be relate to urban or spatial process. You can augment the selected dataset using other datasets (demographic, social, economic, environmental).

- London crime rates (https://data.london.gov.uk/dataset/recorded_crime_summary)
- Prevalence of Childhood Obesity, Borough, Ward and MSOA (https://data.london.gov.uk/dataset/prevalence-childhood-obesity-borough)
- Road safety data in UK (https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-accidents-safety-data): this dataset contains information about road accidents in the UK, including the location, severity, and relevant attributes such as weather, road conditions, and vehicle types.

The data analysis process should be captured within a **single Python notebook**. This notebook should contain all of the code used to complete each of the three stages of the work, in addition to the full documentation of the analysis process and interpretation of results. The documentation must be a **maximum of 1500 words**; the Python scripts and comments are not included in this word limit.

In terms of *how many methods should be used*, you can use up to four methods that are suitable for the research question. If you use a method incorrectly (e.g. using k-means for regression), you will be penalised.

A breakdown of how the notebook will be marked is as follows:

- Analysis context and research questions – 15%
- Data collection, handling, and presentation – 15%
- Correctness, depth, and scope of data analysis – 35%
- Visualisation – 10%
- Quality of writing – 15%
- Creativity of analytical work – 10%

Please consider the following issue in your submission:

1. Should you use “rate” or “count” in your research? In most cases, “rate” is more appropriate than “count”, because a rate normalises the data by accounting for population size or scale. For example, in road safety research, the rate of traffic incidents is calculated by normalising the count of incidents by the traffic volume or population size.
2. Are the references both relevant to the topic and credible (e.g. they are sourced from Google Scholar or CrossRef)? Using unverified references generated by LLM is BAD ACADEMIC PRACTICE and will be penalised.

3. Are the methods used directly relevant to the research question? The inclusion of irrelevant or an excessive number of methods can reduce the quality of the submission.

In addition, a submission at the distinction level should not have any of the following major problems listed below (this is not an exhaustive list of major problems):

1. Including identification columns (aka ID column) for analysis, unless this is well motivated;
2. Incorrectly treating categorical variable as numerical ones;
3. Conducting clustering analysis on dataset with only one variable;
4. Conducting principal component analysis on dataset with two variables and using two principal components for further analysis;
5. Inconsistency between code results and discussion, e.g. a random forest model achieved 80% accuracy but the discussion stated that this accuracy is 90%.
6. The code in the Python notebook is not run or contains errors;
7. The Python notebook and the pdf file are largely different;
8. Using too many methods (more than five) in the analysis and the methods are not related;

Submission

The submission consists of two parts: Part 1: a Python notebook (or a zip file containing the Python notebook and relevant dataset files); Part 2: a PDF file that is exported from the Python notebook in Part 1. Please submit these two parts separately in two tabs on Moodle. The submission timestamp for your submission is determined based on the latter of the two parts. The following situations will lead to mark of 0: failure of submitting either Part 1 or Part 2; the content of the PDF file in Part 2 is not consistent with the Python notebook in Part 1.

At submission, **the notebook should be able to be fully executed within several hours.** Please share the dataset in a Github repo and then remotely read this dataset in the notebook (e.g. using 'read_csv' function as shown in workshops). If the data size exceeds the file size limit of Github (100 M), you could submit a .zip file containing the notebook and data file. Regarding libraries, please stick to the libraries within the recommended and original computing environment (via docker/Vagrant/Anaconda). If you really need to use other libraries (including fastai), you would need to clearly state the names and version numbers of these libraries.

If the data cleaning and pre-processing stages require considerable time for execution, it is adequate that the processed data is provided, alongside a detailed description of the processing phase. The assessors will return work that has not been provided in an easily executed format, which will suffer late penalty deductions.

Before submission, please use the Jupyter function of 'Restart & Rerun all' to ensure that the codes are viable and results are well presented. Penalty will apply if the code is not run or the results contain errors or the results are not clearly presented. Then, please save the executed Python notebook as a PDF file. To do this, in the web browser that runs the Python notebook, you can right click on the browser, select 'Print', then select 'Save as PDF' in the printer dialog box and then select the folder to store the PDF file. Other ways to export the PDF file are acceptable. Note that the PDF file should be text-selectable.

If you get the following warning after submitting the Python notebook or zip file to Moodle, you can safely ignore this warning.

You must upload a supported file type for this assignment. Accepted file types are; .doc, .docx, .ppt, .pptx, .pps, .ppsx, .pdf, .txt, .htm, .html, .hwp, .odt, .wpd, .ps and .rtf

Structure of the notebook

Please use the Python notebook Template_submission_CASA0006.ipynb to format your submission and remove the instruction sentences before you submit.

The following sections should be included in this notebook:

- Introduction
- Research questions
- Data
- Methodology
- Results and discussion
- Conclusion
- References

In the Introduction, you need to include at least three relevant studies. In ‘Research question’, you need to explicitly state the question ending with a question mark. For example, ‘what is the relationship between Covid-19 mortality rate and local deprivation in the UK?’ or ‘Is it possible to predict Covid-19 mortality rate using socio-demographic variables in the UK?’

A title of the notebook is needed.

Other Datasets

Below is a list of interesting datasets, which can be used to augment the data you have selected. You can combine these datasets with a wide range of methods, including making predictions, obtaining data groups, or causal analysis.

1. UK census data: this website contains UK census data - a wide range of data to play with.
2. Cycling datasets in London
    a. Location of London Cycle Hire Scheme
    b. TfL Cycling data: this website contains a wide range of cycling data, including cycle parking, trips of Santander bike hire, etc.
3. Cycle Flows on the TFL Road Network: it contains an index that is used to represent increases in cycle flows on the TfL Road Network (TLRN) over time. It does not represent the total number of cyclists in London. Automatic cycling counters are pieces of monitoring equipment that emit a magnetic field that detects the presence of a moving cycle.