CASA0006: Data Science for Spatial Systems
Assessment Guidelines

Deadline 5pm, 22nd April 2024, Monday, UK Time
Word Count Maximum 2000 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

Select any open dataset relating to an urban or spatial system of your choice and conduct an advanced analysis of the dataset. A complete data analysis process should be undertaken – this will include **data validation and cleaning**, a **data pre-processing** phase (e.g. text, image, clustering analysis), and **comprehensive analysis** (including relevant visualisations) of the data, identifying important trends and insights contained within the dataset. Each stage of the data treatment and analysis process should be well documented and keeping with the exploratory, narrative theme described during the course. Marks will be awarded for both the technical analysis process and the interpretation and choice of analysis methods. The dataset (or datasets) you choose to analyse is left completely open and should relate to an urban or spatial process.

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 2000 words**; the Python scripts and comments are not included in this word limit.

In terms of 'how many methods to use', you are not supposed to use all methods taught in the module. Rather, you can use two 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. Please choose the methods carefully and do not use more than four methods.

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

- Analysis and interpretation of data – 70%
  - Analysis context and aims (incl. reference to relevant literature and projects)
  - Data collection, handling, cleaning and management
  - Depth and scope of data analysis
  - Appropriateness of data visualisation
  - Interpretation and reporting of analysis and major findings
  - Clarity of presentation of results
- Demonstration of technical skills – 20%
  - Choice and rationale of data analysis methods used
- Creativity of analytical work – 10%

Submission

The submission consists of two parts: Part 1: a Python notebook (or a zip file containing the Python notebook and other 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 submission 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 quickly.** 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 satisfactory that the processed data is provided, alongside a detailed description of the processing phase. If you use SQL to pre-process the data, please provide the processed data without including the details of SQL. The assessors will return work that has not been provided in an easily executed format, which will suffer late penalty deductions.

Before your submission, please use the Jupyter function of 'Restart & Rerun all' (or equivalent functions) to ensure that the codes are viable and results are well presented. Penalty will apply if the code is not run or the results are not clearly presented. 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. You can use other ways to export the PDF file. 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

These sections should be included in this notebook:

- Introduction
- Literature review
- Research question
- Presentation of data
- Methodology
- Results
- Discussion
- Conclusion

You can combine 'Introduction' and 'Literature review' into one section of 'Introduction', or 'Results' and 'Discussion' into a section of 'Results and Discussion'. In the literature review, 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. You can use the proposed research question as the title, and other options are acceptable.

Example Workbooks

Listed below are a few example data analysis projects using Python and various libraries, combining code and narrative (to varying extents) within a notebook format. In general, we expect a **more systematic and complete analysis than that offered here** – following the steps outlines above.

- San Francisco Drug Geography - http://nbviewer.jupyter.org/github/lmart999/GIS/blob/master/SF_GIS_Crime.ipynb
- New York Taxi Analysis - https://anaconda.org/jbednar/nyc_taxi/notebook - Excellent visualisations
- Graph Properties of the Twitter Stream - http://nbviewer.jupyter.org/gist/fperez/5681541/TwitterGraphs.ipynb
- Clustering Samsung smartphone accelerometer data - http://nbviewer.jupyter.org/github/herrfz/dataanalysis/blob/master/week4/clustering_example.ipynb
- Exploratory Analysis of the 2014 World Cup Final - http://nbviewer.jupyter.org/github/rjtavares/football-crunching/blob/master/notebooks/an%20exploratory%20data%20analysis%20of%20the%20world%20cup%20final.ipynb
- Data mining Twitter using tweepy - http://nbviewer.jupyter.org/github/hugadams/twitter_play/blob/master/tweepy_tutorial.ipynb?utm_content=14023248&utm_medium=social&utm_source=twitter - very informative!

Once marked, we would encourage you to submit your completed workbooks to nbviewer.jupyter.org or anaconda.org for wider sharing.

Examples Datasets

We would encourage you to find an interesting dataset that you all want to work on. Here are a few examples in case you are struggling to find one.

- NYC GPS taxi data - http://chriswhong.com/open-data/foil_nyc_taxi
- Yelp dataset - https://www.yelp.com/dataset
- UK Land Registry house sales data - http://landregistry.data.gov.uk
- Stop and Search Data by US State - https://openpolicing.stanford.edu/data/
- Traffic Accident and Traffic Flow data for 16 years - https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales/settings
- Real-time crime data in Seattle - https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp
- Various FOI data releases can be found on WhatDoTheyKnow - https://www.whatdotheyknow.com/list/successful
- Crime Data in Buenos Aires - https://github.com/ramadis/delitos-caba
- Lots of open data for Bahrain - https://datasource.kapsarc.org/pages/home/
- City Cellular Traffic Map - https://github.com/caesar0301/city-cellular-traffic-map
- Beijing GPS taxi data - http://research.microsoft.com/apps/pubs/?id=152883
- International Migration data - http://www.global-migration.info/
- Plant Diversity in American National Parks Biodiversity - https://www.kaggle.com/nationalparkservice/park-biodiversity/data
- Wildlife Trade Database - https://www.kaggle.com/residentmario/cites-wildlife-trade-database/data
- H1-B Visa Petitions - https://www.kaggle.com/nsharan/h-1b-visa/data
- Baltimore Crime Data - https://www.kaggle.com/sohier/crime-in-baltimore
- Chicago Crime Data - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- AWS Honeypot Cyber Attack Data (with originating latitude/longitude) - https://www.kaggle.com/casimian2000/aws-honeypot-attack-data/data
- Vancouver Crime Data - http://data.vancouver.ca/datacatalogue/crime-data.htm

Below is another list of interesting and more recent datasets, which can be used to construct more complicated datasets and to answer relevant research questions. You can combine these datasets with a wide range of methods, including making predictions, obtaining data groups, or causal analysis.

1. 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.
2. 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.
3. Road safety data in UK
    a. This website contains information about road accidents in the UK, including the location, severity, and contributing factors such as weather, road conditions, and vehicle types.
4. Covid-19 related data
    a. Historical Vaccination: the Weekly COVID-19 Vaccinations data are split by MSOA, so it is possible to link the MSOA-level vaccinations with socio-demographic data.
    b. Covid-19 data: including number of case, vaccination, etc. You can select different area unit.
5. UK census data: this website contains UK census data - a wide range of data to play with.

6. Oyster card data
    a. Unfortunately, the individual or aggregated oyster usage data are not publicly available. If you are really interested in Oyster card data, you can use [Freedom of Information requests](https://tfl.gov.uk/corporate/transparency/freedom-of-information) to ask for data. Note that it takes time to get a response.
7. New York City Taxi Trip Data
    a. This dataset contains information about taxi trips in New York City, including the pick-up and drop-off location, fare amount, and tip amount.
8. Citi Bike Trip Data in New York
    a. This dataset contains Citi Bike trip data in New York City.