CASA0006: Data Science for Spatial Systems
Assessment Guidelines

Deadline 5pm, 25th April 2022, Monday, UK Time
Word Count Minimum 2000 words (not including Python scripts)

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 Jupyter 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 **minimum of 2000 words**; note that the provided Python scripting is not included in this word limit.

Note that the submission should contain a Python notebook ending with '.ipynb' and probably a data file. Other submitted files will be neglected. For instance, if you submit only a PDF file, you will get a mark of 0.

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 relevant to 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 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%

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.

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'. Note that 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, but other options are acceptable.

Example Workbooks

Listed below are a number of 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.

- Using Python to see how the *Times* writes about men and women - http://nbviewer.jupyter.org/gist/nealcaren/5105037
- How Clean are San Francisco's Restaurants? - http://nbviewer.jupyter.org/github/Jay-Oh-eN/happy-healthy-hungry/blob/master/h3.ipynb
- Predicting use on NYC Metro - http://nbviewer.jupyter.org/url/www.asimihsan.com/articles/Intro%20to%20Data%20Science%20-%20Final%20Project.ipynb
- 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
- Buzzfeed analysis of Segregation in St Louis - http://nbviewer.jupyter.org/github/buzzfeednews/2014-08-st-louis-county-segregation/blob/master/notebooks/segregation-analysis.ipynb - needs better documentation!
- Graph Properties of the Twitter Stream - http://nbviewer.jupyter.org/gist/fperez/5681541/TwitterGraphs.ipynb
- Logistic models of well switching in Bangladesh - http://nbviewer.jupyter.org/github/carljv/Will_it_Python/blob/master/ARM/ch5/arsenic_wells_switching.ipynb - lacks descriptions of the data
- 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!
- Flight Arrivals - http://nbviewer.jupyter.org/github/ResearchComputing/Meetup-Fall-2013/blob/master/python/lecture_27_arrival.ipynb - lacks full documentation!
- Very nice analysis of how the Circle Line rogue train was caught with data - https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a#.oabdxcg86 - GitHub notebook, rather than Jupyter

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

Examples Datasets

We'd 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
- Flight data (requires Google account) - https://bigquery.cloud.google.com/table/bigquery-samples:airline_ontime_data.flights
- 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 lat/lngs) - https://www.kaggle.com/casimian2000/aws-honeypot-attack-data/data
- Vancouver Crime Data - http://data.vancouver.ca/datacatalogue/crime-data.htm