UNIVERSITY OF LONDON

MSc Data Science

Module: Data Programming in Python DSM020-2021-APR

Coursework: April 2021

- Please Note: You are permitted to upload your Coursework in the final submission area as many times as you like before the deadline.

- If you upload the wrong version of your Coursework, you are able to upload the correct version of your Coursework via the same submission area. You simply need to click on the 'submit paper' button again and submit your new version before the deadline.

  In doing so, this will delete the previous version which you submitted and your new updated version will replace it.

- Please note, when the due date is reached, the version you have submitted last, will be considered as your final submission and it will be the version that is marked.

- **Once the due date has passed, it will not be possible for you to upload a different version of your assessment. Therefore, you must ensure you have submitted the correct version of your assessment which you wish to be marked, by the due date.**

Coursework Description

This assignment is worth 70% of the total grade for the module. This will involve producing a project that you pursue in the second half of the module. In this coursework, you are going to build on the exploratory data analysis work you carried out in the midterm coursework. You are going to carry out an in-depth analysis of a dataset of your choosing, using appropriate data processing, analysis and presentation techniques.

The main objectives for this coursework are to:
- state clearly your aims and objectives in carrying out this analysis.
- use data processing techniques to prepare and process the data into useful, analytical form.
- use statistical and other appropriate metrics to analyse and evaluate the data.
- use visualisation, tabulation or other techniques to summarise and present the data.
- present your work as a report in the form of a single Jupyter notebook file using markdown, python code and other appropriate formats for presentation.
- develop an evidence-based narrative in the report.
- use error handling techniques to ensure your data processing pipeline is robust.

Plagiarism:

This is cheating. Do not be tempted and certainly do not succumb to temptation. Plagiarised copies are invariably rooted out and severe penalties apply. All assignment submissions are electronically tested for plagiarism.

Deliverables

Your report should be submitted as a **single** Jupyter Notebook. This notebook should include all acquisition steps, pre-processing and any changes to the data that are deemed appropriate. There should be a clear design rhetoric **throughout** describing the different challenges and conditions. Your approach should be **descriptive, analytical** and facilitate **technical merit.**

Your brief is to explore a manageable data science project, analyse and evaluate the necessary dataset in a usable form. You will need to submit your notebook as well as any resources that you have used throughout the exercise.

Please note:
• Visualisations can be presented inline in the Notebook, or in separately exported files for instance where a graph or diagram is too large (e.g. PNGs.)
• You should include a working sample of your dataset, not exceeding 10MB.
• You should include a requirements.txt file plus any additional instructions about how to replicate your approach and outcomes.

The marking criteria below define the expected outcomes from the project. For further information about expectations at this level of study please refer to Appendix C – Assessment Criteria in the Programme Regulations.

	Part 2	Marks awarded	
a	Aims and objectives are clearly defined, measurable and realistic	10	Your **aims** and **objectives** should clearly identify key research themes for exploration that are: Simple[2] Measurable[2] Achievable[2] Realistic[2] Timescaled[2] These should relate to the outcomes of your project in some way e.g. showing that you achieved each aim and objective in the evaluation/conclusion section of your report.
b	Robust error checking and handling procedures	10	All code cells should run without issue or requiring refinement [2]. You should describe cases for concern or verification. [2] For instance in creation of files where overwrites might occur you should signpost this (ideally using markdown.) You should introduce error handling scenarios where a procedure might be particularly problematic or some process of verification is necessary. [2] You should attempt to utilise test-driven development for some portion of your code and describe this function with a clear annotation. [2] Error handling should be fit-for-purpose and shows that you have handled these issues sufficiently. [2] For example, for web scraping you should be handling 404 and similar types of error codes.
c	Data capture, processing and analysis	10	There should be evidence of techniques used in each case to capture, process and analyse data: - You should systematically describe your process of data exploration to determine that the data is fit for purpose [2] - You should show evidence of data validation [2] For example, this could involve comparing each column/row against some given expectations and setting threshold values to identify outliers or under/over performers. - Identification of trends in the data to suggest further exploration of concepts. [2] For example you might wish to explore particular timeframes, groups or categories for exploration. - You should show that data is being manipulated in a way to converge on your research goals and improve on the utility of said data. [2]

			For example you may wish to engage in validation of a static dataset using web scraping and employ data cleansing or sanitisation as processes for further refinement/verification. For higher marks your analysis should involve the utilisation of multiple patterns of exploration [2] e.g. numerical analysis of data, removing NA values and scaling data, leading to the usage of a pretrained model for classification of objects.
d	Simple code and clear commentary	10	Students will be awarded marks for brevity of ideas as well as readability. Clarity of expression is important here. Examples of this should include: • The use of libraries to simplify solving mundane or repetitive processes. [2] • The use of sensible variable names e.g. not myVar or this_data. [2] • Analysis of the technical elements that validate the approach (e.g. removing or aggregating null values.) [2] • Clear headings and/or structure of the document in relation to both technical and analytical components. [2] • Code should be clearly described in terms of how or why it is being used. [2]
e	Data processing pipelines	10	The report should show clear evidence of iterative development of solutions [5] For example this might involve moving comparing boundaries in your data and exploring fringe cases or outliers to identify issues. When passing and manipulating data, then pipelines should be clear and traceable in the context of your code. [5] For example, you might want to create new dataframes each time you move through a new set of objectives or a new research lens for exploration so as not to overwrite and obfuscate previous analytical steps and stages.
f	Evaluation/ conclusion are well formed based on evidence	10	The conclusions should be soundly based on the analysis and relevant to the objectives of the project. [2] e.g. 'I aimed to explore a,b,c and achieved this through the outcomes of x,y,z.' There should be a clear summary of outcomes based on: a) own work. [2] e.g. "I was successful because 'x' is better than 'y' and 'z.'" b) reading/external material. [2] c) critical evaluation of both own work [2] and external references[2]. e.g. referencing two approaches explored in the course and identifying alternative approaches from the core text or other sources would be appropriate.
g	Different tools have been considered	10	There should be clear evidence of comparison of tools selected to achieve a given outcome. [6] There should be a justification for utilising a particular library, toolkit, technique etc. [4] For example: are you using pandas to handle tables and NumPy for dealing with numerical or statistical data?

			are the libraries that you use efficient? accurate? are statistical values calculated differing depending on libraries that you use?
h	Advanced techniques used	10	Examples might include advanced techniques explored in the course such as TF-IDF for NLP analysis, web spiders for scraping, utilising an SQL database for complex queries, pre-made machine learning algorithms etc. This will depend largely on the research question that you choose to explore and will require sound application of techniques for purpose.
i	Project novelty	10	Very significant ability to evaluate critically existing methodologies and suggest new approaches to current research or professional practice. [5] Very high levels of creativity, originality and independent thought. [5]
j	Exceptional work	10	Very significant ability to plan, organise and execute independently a research project, coursework assignment or examination question. [4] Very significant ability to evaluate literature and theory critically and make informed judgements. [4] Outstanding levels of accuracy, technical competence, organisation and expression. [2]