UNIVERSITY OF LONDON

MSc Data Science

Module: Data Programming in Python

Coursework: Exploratory data analysis

Coursework Description

This assignment is worth 30% of the total grade for the module. This will involve producing a proposal for a piece of work that you intent do pursue in the second half of the module.

The assignment is to produce a proposal for a project of your choosing. This includes defining some aims and objectives of your project, acquiring and utilising a range of programming techniques to ensure that your data is suitable for analysis. You will rely on critical, analytical skills to explore your dataset through some exploratory data analysis steps and to identify some of the key challenges of working with said data. This is a chance to produce your first substantial data processing pipeline and prepare your data for analysis for the final coursework assignment of this module. You will also describe the formal approach you have taken, including any design decisions along the way. This might include but is not limited to discussions around how the data was captured/retrieved, some summary statistical values that pertain to the data as well as more refined discussion that helps to form more concrete research questions that you can explore in your second coursework assignment.

You should submit a single Jupyter Notebook and any related scripts or SQL files included in a single ZIP archive. The notebook should contain a description of your approach as well as any/all processing used to manipulate, cleanse and sanitise the data for purpose, visualisations, tables etc. If your dataset exceeds 10MB, then include a working sample of the data that can be used in place of the full dataset.

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.

Example themes - you do not have to select one of these. They are just to give you an idea.

Theme	Scope	Example project
Premier League Football	Projects could focus on a dataset around English Premier League Football. This could include any data from 20 February 1992 up until the current season.	'How to get relegated, an analysis of poorly performing teams in the English Premier League.'
Literary Masterpieces	Projects could explore famous plays, sonnets and poems.	'What's in a name? An investigation into the names and content of the works of Shakespeare.'
HTML and Markup	Projects could focus on exploring markup from one or more of the top 50 websites, according to Alexa.	'An analysis of the semantic features of streaming websites.'

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 design a manageable data science project, and acquire 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.

For this exercise you should:

- Acquire and prepare your dataset. Here, you will be expected to seek out and find your own dataset – presumably online. Be sure you are allowed to share the data with others and to carefully anonymise the data if sensitive information is present.
- Preparation might involve collation and/or manipulation of data into a usable format. It may involve collecting data from multiple sources for completeness or to verify the integrity and accuracy of the data.
- It may involve creating a database or a flat file format to store and manage data, for instance if you are working with very large datasets or performing lots of create, read, update and delete type operations.
- It may involve writing Python which produces a dummy dataset, for instance if you are working with sensitive data. If this is your preferred option, you may need to think carefully about what you would expect the results of an analysis to look like, so that you can generate the data accordingly (i.e. if generating random numbers, choose a function or method which produces a realistic distribution, and perhaps a realistic amount of noise too.)

• Explain what programming techniques you have used in the preparation of your data (including any command-line or SQL programming.)
• Outline the idea behind your project (i.e. context, significance, expected outcomes.)
• Briefly detail what you intend to do with the data. You are not expected to explain the exact techniques you will use, though it is important that you identify your process as you work and any high-level features of the data.
• You should carefully consider any weaknesses or potential caveats in your approach and present these too.

You should submit a report, presented in a single notebook, which should include:
– introduction/context.
– brief description of data set (or output a sample), including relevant information (i.e. how it was obtained.)
– Some high level descriptions of the data, where it has come from and its appropriateness for your path of exploration.
– summary of key findings/insights.
– some form of discussion/critical analysis.
– conclusion and further work.
– references to any resources used.

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.

Rubric element	Part 1	Marks available	Example considerations
a	Data chosen is interesting enough to facilitate some insights	10	Data chosen is of sufficient complexity to demonstrate students' ability to • Code in python. • Manipulate data in python. • Understand conceptually the changes that may need to be made to data in a data science analysis. • Analyse the data in some capacity.
b	Data is relevant to project aims/objectives and use of data source is clearly justified.	10	Data is relevant to the project brief and list of topics. Data source is clearly justified including: • Origin of data described clearly • A good explanation as to why this data source has been selected. • A clearly identifiable case for working with this specific type of data (e.g. column headings relate to research question.) • Format of data is suitable for analysis (e.g. CSV -> dataframe/numerical analysis.)
c	Project background is clearly defined (e.g. use of literature, research or pre-analysis.)	10	Should include a summary as to: • Why the field is of interest/relevant • Limitations of the work • How the student aims to address these limitations (e.g. new research question.) • The dataset(s) in it/their current state and appropriateness for purpose (e.g. analysis.)
d	Dataset has been sufficiently prepared for analysis (e.g. clean, free of errors, null values handled.)	10	• Data set has been processed to remove illegal values, e.g. characters in number fields through regex validation. • Null / missing values have been addressed e.g. comparing a web scraped example against a static dataset for accuracy or producing sufficient variation if you are generating a randomised sample of data. • Some basic checks have been done for out of bound values for numeric and categorical quantities (e.g. removing a field where a person is specified as being fifteen feet tall.) • Dataset is in the correct format for analysis utilising a chosen technique (e.g. word based data has been tokenised, stemmed, lemmatised etc.)
e	Ethics of use of data have been considered	10	• Description of where the data has come from e.g. open or proprietary or a combination of both. • Considerations about usage/reusage of data. Who owns derivative data/analysis. • Consideration around implications of utilising data for purpose (e.g. is there power to discriminate? Could research summaries produce dangerous or harmful assumptions?) • Considerations of the data processing pipeline. Is the data readily accessible in your notebook? Anonymised? Can it clearly be identified what has been done with the data and that there is no potential for personally identifiable distinctions to be made?

f	Clear rhetoric for modifications to data (e.g. converting between formats, replacing null data with aggregate data)	10	a. Data is modified. b. There is a justification that is reasonable for the modifications c. The modifications add value or utility in some capacity (e.g. descriptive power, performance improvement for analysis.) d. Changes to data utilise advanced techniques.
g	Code is clean (not verbose.)	10	• using functions where repetition of processes are necessary. • Commenting is done in line or markdown is used to separate elements • LaTeX, images and such are used to improve clarity of expressions and ideas. • Code is neat and orderly e.g. lambda functions for non-repetitive processes.
h	Code is functional (ie free of errors, reproducibility of results.)	10	a) Be reproducible in the current notebook format. b) Use proper conventions e.g. relative path vs absolute. c) Be explained or described where libraries are used in relation to their utility/ability to solve a particular problem in an efficient manner. d) Runs without performance issues e.g. long wait times. Exceptional circumstances such as complex machine learning processes may be valid in this case though you may wish to do things like reduce number of epochs and summarise outcomes graphically.
i	Data has been captured using some technique (e.g. web scraping, import/export from database.)	10	Students should show evidence of utilising a format which is appropriate for the data set. In terms of acquisition, this might be described as technical steps such as web scraping or the setup and maintenance of a database schema. Equally, students who choose to utilise a dataset as is should justify why the dataset is in the correct format and fit for purpose.
j	Readability of code	10	Notebooks should be: 1. Structured with a logical set of processes procedures 2. Be systematic and rigorous e.g. step 1, step 2, step 3 with clear separation of processes by cells/blocks of code 3. Interspersed with an appropriate level of discussion to show understanding of how and why concepts are used 4. Not overly verbose e.g. with thousands of comments to describe print statements