A Data Programming Project

Now that you have had a chance to explore some techniques and tools in Python, it is time to start working on your own exploratory data analysis project. This is a chance for you to explore a research area of your choosing. You will identify a clear agenda for research and explore this topic at a high level.

Expectations:

- Identify your own research area and questions, including importing knowledge from external sources.
- Acquiring a dataset that is fit for purpose.
- Exploring the dataset through different lenses, identifying key features and potential flaws in the data.
- Produce a systematic, rigorous and well-reasoned report on how you work through the dataset.
- Describe at both a technical and analytical level, how and why you are approaching the problem space in a particular way.
- Identify gaps in your approach, the dataset and any techniques, tools, libraries or data structures that you choose to utilise.
- Consider the ownership (provenance) of data through a data processing pipeline and how this might manifest.
- Consider how data can be prepared, refined and explored for further analysis e.g. for a final year project.
- Critically analyse, evaluate and summarise findings from a mini-research project.
- Reflect on both processes and outcomes of your project, including any missing steps or stages.
- Give a valuable account as to how your analysis provides useful and interesting insights around some dataset.

  You should present your work in a single Jupyter Notebook (.ipynb file) as part of a larger (ZIP) archive of files. Any data that you use should also be included and readily accessible for checking – included in the ZIP archive. Your ZIP archive should not exceed 30MB in total, including your ipynb file and any data that you choose to utilise. The dataset should not be more than 10MB in total size.

  The marking rubric includes a description of expectations and deliverables, where sections a-j are each worth a total of 5 marks.

Sub part	Criteria	Marks awarded	Mark breakdown
a	An introduction to the research space.	5	The report clearly demonstrates students' ability to: • Produce clearly defined aims and objectives for an independent research project. • Acquire a dataset for working with. • Utilise the dataset through an exploratory data analysis in Jupyter Notebook. • Write in a way to communicate ideas and concepts clearly. • Present a clear summary of the area of research chosen.
b	Data is relevant to project aims/objectives and use of data source is clearly justified.	5	Data is relevant to the project brief and list of topics. Data source is clearly justified including: • Origin of data described clearly including data source and acquisition techniques used. • A good explanation as to why this data source is appropriate for the research question posed. • 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.) • A consideration of at least two other datasets and their potential strengths/weaknesses for your chosen research topic.
c	Project background is clearly defined (e.g. use of literature, research or pre-analysis)	5	Should include a summary as to: • Why the field is of interest/relevant • That the topic has not been previously explored and/or research questions have not already been answered. • Scope of work e.g. "I will analyse x and y but not z." • Steps and stages in your analytical data processing pipeline. • A description of how you will evaluate your aims and objectives based on your chosen approach.
d	Dataset has been explored technically	5	• Data set has been processed to remove illegal values, e.g. characters in number fields through regex validation. • Data is in the correct format for analysis e.g. numpy nd array, dataframe, with a

			clear distinction as to why this format is correct and appropriate. • Checks have been done for out of bound values or for numeric and categorical quantities (e.g. finding min, max values, sorting data into logical groups for analysis such as top 10%, bottom 10% etc.) These should be justified in terms of the research question specified and the categories defined/described. • Depth of exploration draws out some interesting or valuable insights (e.g. problematic data in higher thresholds, comparing two disparate datasets for accuracy.) • Data is in an appropriate format to carry out further analysis e.g. for a machine learning pipeline or for generating plots, diagrams and charts.
e	Ethics of use of data have been considered	5	• Description of where the data has come from e.g. open or proprietary, licensing and wider considerations around provenance. • Considerations about usage/reusage of data e.g. does the analysis have the potential to create new forms of intellectual property? How is attribution given? • 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? • Any potential biases of the dataset have been considered (e.g. where 80% of the dataset falls into one demographic and 20% for another.)
f	Clear rhetoric for modifications to data	5	• Data is modified (e.g. converting between formats, replacing or removing data, combining multiple datasets for improved accuracy) • There is a justification that is reasonable for the modifications with clear summaries as to how and why the data has been adapted for purpose. • The modifications add value or utility in some capacity (e.g. descriptive power, performance improvement for analysis.) • Data has been changed in a systematic (not arbitrary) way that shows clear understanding and justification in exploration of concepts and ideas. • Changes to data utilise advanced techniques for example by comparing a static CSV dataset against some web scraped data to compare accuracy.

g	Code is clean (not verbose)	5	Examples of clean code might include: • using functions where repetition of processes are necessary or using lambda functions where they are not. • Commenting is done in line or markdown is used to separate elements where there is a clear distinction between code and narrative. • LaTeX, images, tables, markdown and such are used to improve clarity of expressions and ideas. • Code is neat and orderly e.g. easy to read, consistent styles used throughout. • There is a good mix of code and descriptions in the right places e.g. where complex bits of code are present they are neat and well-described.
h	Code is functional	5	Code should: • Be reproducible in the current notebook format including making relevant data sources and libraries accessible and explicit. • Use proper conventions e.g. relative path vs absolute. • Be explained or described where libraries are used in relation to their utility/ability to solve a particular problem in an efficient manner. • Runs without performance issues e.g. long wait times for web scraping tasks (hint: pickle your data or at least store in a flat-file format.) • Handle errors in appropriate places (e.g. for web requests – catching 404s.)
i	Dataset has been analysed	5	Notebooks should clearly evidence: • Some exploratory findings from the dataset, for example through the use of a word cloud and some high-level statistics/descriptive values. • Key features of the dataset that might be problematic e.g. the data is skewed, the dataset contains missing or erroneous data. • A rich discussion drawing insights from the data based on exploration. • Some evaluation and conclusion(s) based on your findings. • A critical evaluation including any flaws in the techniques you have selected.
j	Readability of code	5	Notebooks should be: • Structured with a logical set of processes/procedures including clear, logical headings. • Be systematic and rigorous with clear separation of processes, ideas and manifestations of exploration. • Not overly verbose e.g. including comments to describe print statements.

			• Justified at each stage e.g. “creating a new dataframe here is appropriate so that we do not obfuscate previous steps in our pipeline.” • Consider readability e.g. through removing trailing whitespaces, aligning brackets, avoiding line-wrapping where possible.