

Project Information

Clarifications about the description

The aim of the project is to produce 2 different pieces of data analysis work. Therefore, the 2 datasets you choose should be completely different to each other although they *can* belong to the same domain if you intend to combine them in your analysis. Cases like (but not limited to):

- One dataset being a derivative of the other,
- Both being subsets of the same original dataset,

will inevitably lead to some duplication of parts of the work (literature, objectives, design analysis, code, etc.) for which points cannot be re-awarded.

The datasets are supposed to also be publicly available. There are numerous resources on the WWW with publicly available datasets (see “Sample data sources” section). If you prefer to use a dataset that is not readily publicly available (like data from your workplace), you **must** ensure that:

- You have full permission to use this data in a piece of work that will make both the dataset and analysis thereof public and
- The dataset does **not** contain any personally identifying information.

Basically, by using such a dataset you *will* be making it publicly available so you need to secure the dataset owner’s permission to do so in advance.

Structure and Rating Grid

Objectives and Literature Review

As with every piece of data analysis, you should ideally have a question or set of questions you expect your work to answer; these are your objectives. They will be graded for realism, imagination, ambition and clarity of expression.

Your objectives are inherently tied to the state of knowledge on the domain, which you can gauge via reviewing the domain literature. This should be presented as a synthesis of the referenced works, not a compilation of summaries. Your literature review should be properly (Harvard-style) referenced. It will be graded for quality, depth and extent.

Dataset description

Your chosen datasets should be included in their original form as ancillary files. If they are prohibitively large, you should include a well-chosen, representative subset. Where and how the datasets were located and downloaded should be clearly shown. They will be graded for richness, depth and interest factor.

The dataset description should encompass all columns and detail data types, ranges, special cases, etc. More focus should be given to the columns that will ultimately be used for the data analysis. All interesting and/or pertinent information about the dataset should be presented. The description will be graded for detail, structure, clarity, etc.

Analysis approach

Your data analysis should be designed in advance and the design documented via description and visual aids such as tables, flowcharts, and other appropriate schematics. Please note that screenshots of your code do **not** count as such in the general case and should be avoided unless there is a specific reason why they are appropriate.

There are 3 established approaches to data analysis^[1]:

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- Knowledge Discovery in Databases (KDD)
- Sample, Explore, Modify, Model and Assess (SEMMA)

Of the three, CRISP-DM and KDD are the most generally-implementable, whereas the design of SEMMA assumes the use of the Enterprise Miner software from SAS. You can find more information on them in the “Miscellaneous Resources” section. You are free to adopt any of these for your project or follow your own. **No points will be awarded for following an established methodology.**

Your data analysis process will be graded for robustness, adherence to commonly accepted standards and completeness (inclusion of some kind of testing process, result evaluation, etc.) Description thereof will be graded for detail, clarity, appropriate use of visual aids, as well as the quality and variety of the latter.

Analysis results and presentation

The results of your analysis should go as far as possible towards reaching your prior stated objectives (i.e. answering the questions you were hoping to answer). Note that a robust conclusion that the dataset or the analysis aren't enough to reach a specific conclusion isn't a failure but is, in fact, a positive result!

Your analysis will be graded for robustness, appropriate use of statistical methods, appropriate use of code, etc. The presentation of your analysis will be graded for clarity, depth of information, appropriate use of visual aids (graphs, plots, etc.) and quality thereof.

R code

You are required to use R to an extent that showcases your aptitude with the most important operations learned during the course (file I/O, control structures, functions, etc). A substantial amount of code is expected.

Your code will be graded for extent, quality, good use of coding conventions (comments, variable naming, etc.). Note that the use of libraries, while encouraged, will **not** be given extra marks.

Project report

Your ultimate project report, encompassing all the above, will be graded for structure, presentation and quality of its discussion of challenges. There is no requirement to structure your report as a scientific paper, though you are free to do so if you prefer.

Please **do** include your complete R code as an appendix. Obviously, your code does not count towards the word count requirement.

Deliverables

Your main deliverable will be the project report, which should be submitted via the Turnitin submission form, appropriately entitled “Project report submission”. Ideally it should be in PDF format, but MS-Word and other similar formats are also acceptable. If in doubt, ask.

Your ancillary files should be submitted via the “Project ancillary file submission” form as a zipped archive. It should contain *at least* the following resources:

- Your R code as 1 or more .r file(s),
- Your 2 datasets in the state they are input into your R code.

Please use relative filepaths in your code, such that unzipping the archive allows all code to be run without any need to edit the filepaths or move the datasets around.

Other included files can (but are not required to) be intermediate dataset outputs, graphs/plots/etc as graphics, or whatever other artefacts you deem appropriate.

Referencing and plagiarism

Ideally, do lots of the former and none of the latter. In case it hasn’t already been drilled into your heads (which it should), any type of plagiarism makes NCI very, very sad and the penalties are severe. Please re-familiarise yourselves with the [NCI policy on referencing and plagiarism](#). Please note that re-submitting your own work (either from the same module earlier or a different one) is *still* considered plagiarism.

The fact that we will be dealing with code doesn’t make much difference in the guidelines about plagiarism. While programmers work by re-using each other’s code constantly, in an academic environment you are still expected to cite any code you use that is not strictly your own. Luckily, code comments provide the perfect mechanism for referencing that code’s origins.

- If you’re reusing a whole script or class, you should reference all the appropriate details (URL, original author, date, etc.) in a large comment block at the top of the file.

- If you're reusing a code snippet, you should reference its origin URL and author in inline comments just above the snippet.

Treat any other case appropriately, always with a view to making it as clear as possible that code reuse has taken place. If in any doubt, reference it!

Sample data sources

This is an indicative, not exhaustive list; there are many more sources of public data not listed here which you are free to use.

US Government Open Data Initiative	https://www.data.gov/
EU Open Data Portal	https://data.europa.eu/euodp/en/data/
UK Government Data	https://data.gov.uk/
Irish Government Open Data Portal	https://data.gov.ie/data
World Bank Open Data	https://data.worldbank.org/
OECD Stats	http://stats.oecd.org/
UN Data	https://data.un.org/
Amazon Web Services Public Data	https://aws.amazon.com/datasets/
Google Trends	https://trends.google.com/trends/explore
Google Finance	https://finance.google.com/finance
Google Public Data Explorer	https://www.google.com/publicdata/directory
Yahoo WebScope	https://webscope.sandbox.yahoo.com/
Moody's Analytics	https://www.economy.com/freelunch/
Pew Research Datasets	http://www.pewresearch.org/download-datasets/
UCI Machine Learning Repository	http://archive.ics.uci.edu/ml/index.php
DataHub	http://datahub.io/search
Quandl	https://www.quandl.com/
NASA Earth Observatory	https://neo.sci.gsfc.nasa.gov/

References

1. Azevedo, A., Santos, M. F. (2008); [KDD, SEMMA and CRISP-DM: a parallel overview](#). Proceedings of the IADIS European Conference on Data Mining 2008, pp 182–185