**Structure:**

1. Objectives and Literature Review
2. Dataset description
3. Analysis approach
4. Analysis results and presentation
5. R code

**Structure – With 5 main topics**

**Objectives and Literature Review**

* Description of the objective(s) of the analysis with reference to basic domain literature to explain the domain purpose of the analyses

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**

* Description of the underlying dataset including an assessment of the data types present, with an emphasis on the data that is actually used in the analytical processes

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 where 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**

* Approach to the analysis, aided by visuals such as diagrams, flowcharts and tables where appropriate

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**

* Project report structure, presentation and discussion of challenges.

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**

* R code demonstrating at least four unique insights. R scripts will be executed as part of the assessment process. It is expected that scripts are fully working, efficient, commented clearly, and do not contain excess 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.

**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 or MS-Word.

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.

**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.

**Referencing and plagiarism**

* 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.

You are expected to cite any code you use that is not strictly your own.