DATA1002/1902 (Sem2, 2019) Project Stage 2

Due: 11:59pm on Sunday November 3, 2019 (week 12)
Value: For data1002: 15% of the unit. For data1902: 20% of the unit

This stage is usually done with the same group members as you worked with for Stage 1. Membership changes will only be made following the process described below. Note however that a new Canvas group has been created for this stage of the project (so that any changes made now, do not affect the marking of stage 1).

Process to divide an existing group: If for any reason any members in the group want to leave (and form a new group which is a subset of the original, leaving the others behind) then they should urgently email the unit coordinator alan.fekete@sydney.edu.au and cc all the existing group members, describing clearly what new membership structure they desire. Please note that a change is not necessarily granted; the decision will be made by the unit coordinator after trying to consult all the people involved.

Process for someone currently alone, to join a group: a member of the existing group needs to email the unit coordinator alan.fekete@sydney.edu.au and cc all the existing members and also the new one, indicating that all the existing members agree to accept the new member. As long as the number of members remains four or fewer, this change will be allowed.

Process for someone to move from one existing group to another: this needs to happen in two stages. First the person needs to request to leave their current group (as described above). If that is approved, they can then follow the process to join a new group.

Problems during the stage: If, during the course of the assignment work, there is a dispute among group members that you can't resolve, or that will impact your group's capacity to complete the task well, you need to inform the unit coordinator. Make sure that your email names the group, and is explicit about the difficulty; also make sure this email is copied to all the members of the group. We need to know about problems in time to help fix them, so set early deadlines for group members, and deal with non-performance promptly (don't

wait till the final week when the work is due, to complain that someone is not doing their share).
If necessary, the coordinator will split a group, and leave anyone who doesn't participate effectively in a group by themself.

The project work for this stage:

For this task, you need to do some interesting analysis on a data set, and to produce (one or more) automated predictive models using machine learning techniques. We expect that most groups will use the data set you already worked with in Stage 1.

If you want to do the extra work to find a different data set, and clean it, you are allowed to do so. Alternatively, you may request a data set from us, and we will supply one that has been cleaned (but it may not interest you particularly).

Advice: Please do not try and divide the work among the group members by type of task. It would be a very bad learning experience if one person does the coding of the analysis, someone else does the coding of the predictive model, and one person only writes the report! Instead, each member should make sure they do some of all the kinds of work: write some analysis code, build a predictive model, write some of the report.

What to submit, and how:

There are three deliverables in this Stage of the Project. For groups from data1002, and four deliverables for groups from data1902. ***Each should be submitted by only one person, on behalf of the whole group.***

Submission 1: Report

Submit a written report on your work, as a PDF document.

- This should be submitted through the link in the Canvas site.
- The report should have three distinct parts for data1002 groups, and four parts for data1902.

Report Part One:

Aimed at a general audience that is interested in the domain *(for example, if your data set is about pulsars, assume the readers are like those of a popular science article on pulsars).*

In this part, you should focus on the **insight about the domain that was gained** from the analysis:

- describe the domain situation,
- identify one or more questions that you are looking to explore, concerning possible relationships between some of the aspects or features of the domain
- the origin of the data you used, and then
- present what your analysis has revealed about the domain (in particular, what you have discovered about the questions you were exploring.
- You should include well-chosen tables and visual displays of the summarised data, along with associated textual discussion.

Report Part Two:

Aimed at people with interest in IT approaches to data analysis *(such as other students in data1002!);*

In this part, you should explain **how** you did the analysis (what tools you used both for analysis and for generating the tables and figures in your presentation)

- you should include the code (or at least all the key parts). If you used Excel to generate some of the figures, describe how you did that.
- It should also explain **why** you did things this way, including
    - things you tried that did not work out (or that you were later able to improve), and
    - what you learned from those earlier less-successful attempts.

Report Part Three:

Aimed at people with interest in IT approaches to data analysis *(such as other students in data1002!);*

In this part, you should explain **how** you produced one or more predictive models (what tools you used to generate the model, and any relevant settings, including what training data you provided)

- you should include the code (or at least all the key parts).
- You should discuss what you have learned about machine learning from this experience. In particular, if you have tried several different approaches (or different configuration settings), you

should discuss their strengths and weaknesses as evidenced by your experience (not simply general comments about them, as found in lectures etc)

Report Part Four: (data1902 groups only)

Aimed at people with interest in sophisticated IT approaches to data analysis *(such as other students in data1902!).*

In this part, you should describe how you have produced an interactive visualization of (some aspects of) the dataset.

- you should include the code that generates the interactive visualization.
- It should also explain **why** you did things this way, especially, why you chose the particular style of interactions to offer the audience
    - Mention things you tried that did not work out (or that you were later able to improve), and
    - what you learned from those earlier less-successful attempts.

Submission 2: Your Source Code

Submit a copy of the source code that you wrote to perform the analysis you have done. **This should be submitted through the link in Canvas, as a single file.**

In most cases, we expect you to submit a Python program (or Jupyter notebook with code). This would include calculations that produced the tables and figures, and also the calculation that generated a predictive model. Make sure the code is easy for a reader to understand. If you used Excel to produce some of the charts, then submit a compressed (archived) directory that contains both a spreadsheet and also the Python source for the other aspects of the processing and predictive model.

Submission 3: Your Clean Data

Submit a single file with the clean data that you used. This could be a csv file or json etc, but If your data was spread among several files, you need to compress/archive them into a single file and submit that.

Submission 4 (data1902 only): Your Interactive visualisation

Submit a single file with one or more html pages that present your visualisation. If you have multiple pages forming a website, please submit an archive that contains all of the pages compressed together. Please name the initial root page as project2group<groupnumber>.html.

Marking:

Here is the mark scheme for this assignment. The marker's evaluation will be made principally on the basis of your report; the submitted code and data may be considered as evidence to check or clarify statements made in the report. Note that all members of the group receive the same mark.

Scale of the data: it is required that your cleaned data set contain at least 100 values (as defined in stage 1, so for a rectangular table dataset, the number of rows multiplied by number of columns must be at least 100)

Outcome of analysis: 3 Marks

(as described in Part 1 of the report).

Note: you will not be penalized in marks if you explore a reasonable question about the domain, by looking at appropriate relationships between some aspects, and then conclude that there is no clear relationship revealed.

- A pass (adequate) score indicates that your report delivers an analysis that explores the relationship between at least two aspects or attributes of the data that are relevant to the question you identified
    - The phrase "explore the relationship" could mean seeing if there is a trend that describes how one attribute's value is influenced by the values of other attributes,
    - or it could mean deciding whether the distribution of values of one attribute is different among different subsets of the data, defined by the values of other attributes, etc.
- A distinction level score (good work) is awarded if,
    - your analysis explores connections that (among them) involve at least three aspects or attributes of the data that

are relevant to the question
- Full marks (excellent work) indicates that you have
    - carried out a sensible exploration of how at least four attributes are related together [that is, you haven't just considered pairwise relationships among the four, but really a four-way relationship].

Methodology of Analysis: 3 Marks

The way you carried out the analysis (as described in Part 2 of the report, and evidenced in the submitted data files and processing).

- A pass score is awarded if your processing correctly produces some meaningful outputs
    - the calculations must relate to the question that you are exploring
    - it is clear that the calculations are correct (either it is internally self-evident, or well explained in comments and the report).
- A distinction score is given if you reach the pass level, and also
    - your analysis deals with at least three aspects/features of the data, and
    - your processing is very well automated in Python, so the whole analysis and table and chart production can be redone for changed data sets with only a command or two).
- Full marks would be awarded for doing the above, and also
    - your analysis deals with how at least four attributes are related together [that is, you haven't just considered pairwise relationships among the four, but really a four-way relationship], and
    - your analysis uses techniques of Python that are more powerful than those taught in DATA1002 (such as more sophisticated libraries, charting functions, etc).

Communication of Analysis Outcome: 3 Marks

The way you communicate the results of your analysis (as shown in Part 1 of the report).

- A pass score indicates that the intended audience could gain knowledge of some aspect of the data, without excessive effort or confusion
    - (as part of this, you need to include some visual

presentations that are helpful; the report should also explicitly point in text to the properties the readers should observe in the presentations).

- A distinction score indicates that the report is well-targeted to make it *easy* for the intended audience to gain *understanding* of what the relationship between aspects of the data, reveals for the question being addressed
    - this includes clearly linking your writing to the audience's background and aims;
    - it also requires that the charts draw attention to important properties of that relationship; as well, the writing must provide a convincing justification of any claims made about the data.
- Full marks is awarded for a report that meets all the Distinction criteria, and
    - the charts and tables provide an extensive understanding of the complexity in the dataset and all aspects that are relevant to the question being addressed
    - the charts and tables attract attention and excite the reader
    - also it leaves the readers clearly having learned something, and clear about what further questions have not been resolved.

Communication of techniques: 3 Marks

The way you communicate the techniques used for doing your analysis and charting (as shown in Part 2 of the report).

- A pass score indicates that the intended audience could gain knowledge of what you did, without excessive effort or confusion
    - (as part of this, you need to include clear descriptions of the computations).
- A distinction score indicates that the report is well-targeted to make it *easy* for the intended audience to gain *understanding* of the techniques you used and of the lessons you learned about the techniques
    - (this includes clearly linking your writing to the audience's background and aims; it also requires that the report reflect on strengths and limitations of the techniques used).
    - It also requires good code style
- Full marks is awarded for a report that
    - meets all the Distinction criteria, and also it
    - conveys a sophisticated understanding of some analysis

technique or code library, that goes beyond what was taught in data1002.

Predictive model: 3 Marks

The predictive model(s) you have produced (as shown in Part 3 of the report).

- A pass score indicates that you have correctly used Python (perhaps a Python library such a scikit-learn) to build a reasonable predictive model for some feature in the dataset, based on values of other features.
    - as part of this, you need to include clear descriptions of the approach you used.
- A distinction score indicates that you have correctly used two different techniques to build predictive models for the same feature (this could be using quite different techniques, or just different hyper-parameter or other settings), that each is explained clearly, and that you have done a reasonable comparison of how the two techniques worked in this situation
- Full marks is awarded for a report that
    - meets all the Distinction criteria, and also it
    - used some machine learning technique that goes beyond what was taught in data1002.

Interactive visualisation: 3 Marks (data1902 only)

The interactive visualisation(s) you have produced (as seen in the submitted web pages).

- A pass score indicates that you have produced an interactive visualization that has some features that are more useful and informative for the viewer, compared to a static visualisation.
- A distinction score indicates that you have achieved the pass criterion, and also, your visualization can be used to explore user-chosen subsets of the data and connections between different attributes. It should support the
- Full marks is awarded for a visualization that meets all the Distinction criteria, and also it is easy ("natural") for the viewer, working smoothly without needing detailed instruction. It should adhere to the "8 Golden Rules" provided by Ben Schneiderman http://www.cs.umd.edu/~ben/goldenrules.html and in particular, it should follow the Schneiderman mantra: overview first, zoom and filter, then details-on-demand (see

https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf)

Report on interactive visualisation: 2 Marks (data1902 only)

The explanation of the interactive visualisation (as shown in Part 4 of the report).

- A pass score indicates that you have correctly used Python (perhaps a Python library such a bokeh) to build a reasonable interactive visualisation.
- A distinction score indicates, as well as the Pass criterion, that you have explained clearly what you did (including having clear well-structured and commented code).
- Full marks is awarded for a report that
    - meets all the Distinction criteria, and also it
    - describes and justifies choices you made in designing the interactive visualization (using the concepts from Schneiderman's Golden Rules and Schneiderman's mantra)
    - reflects sensibly on strengths and weaknesses of the approach you took, and the Python (or Python library) features you used

Advice:

During the project, you need to manage the work among the group members. We recommend that you do NOT allocate a different kind of work to each person. That is, don't get one member to write code, another to produce graphs, another to write text, etc.

Instead, we recommend that every person do each activity (perhaps for exploring the relationships of a different group of attributes). This will be important for preparing each member for the final exam.

Late work:

As announced in CUSP: Late work (without approved special consideration or arrangements) suffers a penalty of 5% of the available marks, on each calendar day after the due date.

No late work will be accepted more than 10 calendar days after the due date.