STAT828 Data Mining – Directed Knowledge Discovery (Data Mining) Project

Project Plan Due Date: Tuesday, 19th March 2019, 9:00am

Report Due Date: Tuesday, 28th May 2019, 9:00am

Poster Due Date: Tuesday, 4th June 2019, 9:00am

This is a group project for internal students but it is an individual project for external student (unless external students want to work in a group and they organize themselves how to work efficiently within a group when you have no chance of meeting face-to-face). External students should let me know if they are forming groups, otherwise I will assume they are working individually. Internal students should let me know if they want to work alone after discussing with me (Ayşe) however my preference for the internal students is to work in groups. The maximum group size is 5 (students).

The data should have both categorical and continuous variables, not a complete data set (missing values would be good so that you learn how to deal with them), and a binary outcome variable. If you have a categorical variable for outcome with more than two categories, you might be able to re-categorise to turn it into a binary outcome. Let me know if you need help with choosing the right data set, but keep in mind that the project plan is designed to assess (for me) the suitability of the data set and whether you understood what needs to be done for the analysis. The plan only be marked when you have the right data set for the analysis, it is going to be an iterative process. If the plan is not sound or the data set is not suitable, you will be asked to re-submit your plan.

Possible sources for data sets for this project:

1. The *Kaggle Data Science Competitions* at https://www.kaggle.com/datasets
2. *Data analysis competitions* of the IASC - International Association for Statistical Computing at http://www.iasc-isi.org/node/276
3. SAS data repository: Data Sets for Teaching and Learning at https://communities.sas.com/docs/DOC-4361
4. Australian Government data.gov.au https://data.gov.au/dataset/mbs-sample-10pct-1984-gz "Linkable de-identified 10% sample of Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS)"
5. The Demographic and Health Surveys (DHS) Program at http://www.dhsprogram.com/
6. Café Data 2.0: New Data From a New and Improved Café (I am not sure this is a suitable data set) http://amstat.tandfonline.com/doi/full/10.1080/10691898.2016.1196064
7. Alzheimer's Disease Neuroimaging Initiative http://adni.loni.usc.edu/
8. UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
9. You can use your own data set related to work, or other studies. If you have your own data[1], you can use it after discussing with me. I do not need to have the data set however you need to provide enough information about the data so that I can judge whether the data set is suitable for this unit. If required I can sign confidentiality agreement.

[1] If the dataset you use is of current commercial interest and has been provided by a commercial organisation, no IBM SPSS Modeler analysis or IBM SPSS Modeler output can be included in any report that may be viewed by representatives of the organisation that has provided the data (other than a representative who happens to be enrolled on the unit and is part of the group analysing the data). However, the report submitted for examination may contain IBM SPSS Modeler analysis and output.

The aims of this project are:

- to apply classification analysis techniques (i.e. decision trees) to create rules for the future;
- to write a professional report of your findings;
- to present your findings in a Poster;
- to assess your own learning;
- to assess your fellow students' posters (details will be provided later); and
- (just for internal students) to work effectively as a team member.

State the Business/Research Problem for your report (e.g. why are you doing this?). Explore the dataset to understand what is in it (descriptive data mining). You can use visualization tools from R and/or IBM SPSS Modeler[1] or other software packages. When you know what is in your dataset, then you can decide which fields (variables) will be used in your data mining explorations, accordingly you can clean and/or prepare your data (i.e. creating concept hierarchies, transforming some of your variables or creating new variables from the ones in the data set) for the predictive data mining. This data exploration should form the description of the data set part of your report. Identify, justify and apply suitable models to your data, and compare the findings of different models to decide which model will be the most useful as a predictive data mining model. Finally, describe how your findings could be used in day-to-day operations of the company. Each report will have (a) project plan, (b) a project report and (c) a poster based on the analysis of the chosen data set.

Apply all the classification techniques you have been taught and compare the various methods. Decide which is (or which group of methods are) the best classification method(s), justifying your choice (i.e. present error rates for prediction and other goodness measurements). You will of course need to use a data set that includes a classification variable (outcome variable). Some methods may not be appropriate anyway, if so state this and explain why. Make sure the training set consists of about 60% of the data that you fine-tune the classifier using pruning etc. on a further 20% of the data (evaluation data) and do the final error rate assessment on the final 20% of the data (test data).

The Data Mining Project Plan Template

Student Name(s) and Student ID(s):

Data Set Name:

Source of the Data (could be URL):

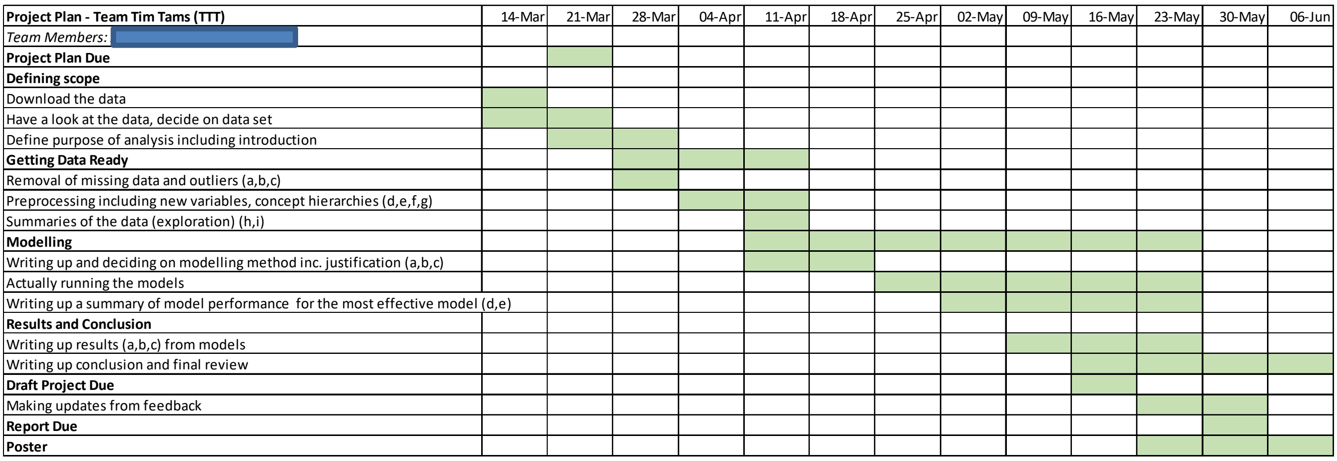
Outcome Variable:

Summary of the Project: (maximum 250 words)

This summary should include:

aim(s) of the project (question(s) to be answered),
why this research is required;
how the expected findings of this report would be useful/beneficial (one or two sentences) to the company and/or society.

Please also include a timeline of your project similar to below:



Project Plan - Team Tim Tams (TTT)	14-Mar	21-Mar	28-Mar	04-Apr	11-Apr	18-Apr	25-Apr	02-May	09-May	16-May	23-May	30-May	06-Jun
Team Members: [redacted]													
Project Plan Due		🟩											
Defining scope													
Download the data	🟩												
Have a look at the data, decide on data set	🟩	🟩											
Define purpose of analysis including introduction		🟩	🟩										
Getting Data Ready			🟩	🟩	🟩								
Removal of missing data and outliers (a,b,c)			🟩										
Preprocessing including new variables, concept hierarchies (d,e,f,g)				🟩	🟩								
Summaries of the data (exploration) (h,i)					🟩								
Modelling					🟩	🟩	🟩	🟩	🟩	🟩	🟩		
Writing up and deciding on modelling method inc. justification (a,b,c)					🟩	🟩							
Actually running the models							🟩	🟩	🟩	🟩	🟩		
Writing up a summary of model performance for the most effective model (d,e)								🟩	🟩	🟩	🟩		
Results and Conclusion													
Writing up results (a,b,c) from models									🟩	🟩	🟩		
Writing up conclusion and final review										🟩	🟩	🟩	🟩
Draft Project Due										🟩			
Making updates from feedback											🟩	🟩	
Report Due												🟩	
Poster											🟩	🟩	🟩

If you have problems related to group dynamics please contact Ayse Bilgin for help. It is better to deal with the problems earlier so that they do not negatively affect your learning.

You might benefit from reading at a short paper regarding “Binary Classification” before choosing your project data set
https://www.cse.iitk.ac.in/users/se367/10/presentation_local/Binary%20Classification.html

The Data Mining Report Template

Use Microsoft Word or a similar open version of word or pdf to create a Data Mining Project Report, containing the following sections (see marking scheme for fine details):

- **Executive Summary or Abstract:** Give the big picture first and then in two or three sentences summarise the essence of what you have done *and discovered* in no more than 15 lines (approximately 250 words). It is a bit like telling the punch line of a joke before telling the joke, but busy executives insist on it these days – they don't want to be distracted by all the whys and wherefores, they just want to "cut to the chase" and get a distillation of what has been achieved.
- **Introduction**: This section provides an overview of your entire document.
  - **Set the scene:** Explain the context in which your project has been developed, i.e. the big picture. Describe the possible business environment and identify a problem (or problems) in current time. Identify intended users for the results.
  - **Project Goals:** Describe the problem(s) to be solved, as a set of general questions which can be answered by the data mining analysis. (Imagine yourselves being the manager of the company that its data being analysed while you are writing the report.)
- **Description of the data set:** This section is for data understanding and descriptive data mining.
  - **Original Data:** List the number of attributes (variables), number of observations (rows), and format of the data, followed by a description of your data set (i.e. graphs, tables).
  - **Data preprocessing:** Explain how you preprocessed your data. Did you find any outliers, if so what did you do? How about missing values? Did you create any new variable(s) from existing variables, if so, how and why? Did you exclude any variables or observations?
- **Methodology:** This section gives the details of modeling methods.
  - Briefly describe the modeling methods you have used. Explain the reasons for using a specific method and not any other.
  - You also need to explain the reasons for choosing a particular modeling option for each specific model and then present the prediction accuracy (or class specific error rates) of running the chosen models (with training, evaluation and test data sets).
  - Provide the comparisons of the models applied to the data and identify the best model or models for presentation of the results. Explain why the model(s) is the best? **(Data Mining Evaluation)**
- **Results:** Display and explain the results for the best model, if possible, in layman terms. For the questions identified above (this could be just one question), state the knowledge you have discovered that provides insights into this question(s). State how well the knowledge answers the question(s), what is missing or requires further analysis.
- **Conclusion:** Provide an ending to this document with a mention of how your findings could be used in day-to-day operations and identify possible extensions for the future. In addition, identify the limitations of your data set and analysis, if there were any. Your conclusion may have quite a lot in common with the "Executive Summary" but it should be more detailed.
- **Appendix:** Computer outputs for all the models applied should be presented here. In addition, anything that is important to present but will take too much space in the main report should be placed in the appendix. Make sure you refer to everything in the appendix in your report.

Note: Your main report should not be more than 10 A4 pages. All relevant graphical displays should be embedded in your report. Font type for the report is Times New Roman. Font size is 12 in the body. Line space is 1 or 1.5.

Note that your poster presentations should cover the above without being too excessive.

What to submit (see the beginning of this document for the dates and times)**:**

The Project Plan is due on Week 4.

1. Upload one project plan per group on iLearn by due date (word document) clearly stating the group members and their student IDs. For external students, one plan for each student, unless they are in a group.

The Project Report is due on Week 12.

1. a. One project report per group (internal)/student (external) on iLearn by due date (word document) clearly stating the group members and their student IDs. For external students, one report for each student, unless they are in a group. The report can have maximum 10 pages.
   b. You can use appendices for extra information that you find important to be part of the report.
   - Font for the report: Times New Roman
   - Font Size: 12 in the body
   - Line space: 1 or 1.5

   c. In addition to group report, each student working in a group will need to have an individual part added to their group report (see next page) which will be no longer than 3 pages.

   Optional: If you want you can submit a printed copy of your project report to Ayşe Bilgin at the beginning of the lecture on Week 12.
2. R codes (as an R script file or text file) and/or IBM SPSS Modeler streams used for your analysis.
3. (**Internal Students**) Each student has to answer the questions listed under "**Your Experience of working in a group for this Project**" and upload her/his answers within iLearn. This is a confidential document and will remain confidential between the lecturer(s) and each student. It will be used for trying to understand the group dynamics and identify the contributions of the students to the project. In addition, it might be used for research on learning and teaching and presented as a conference paper or published as a journal article after making data anonymous.
4. (**Internal Students**) Each student has to fill the "**Peer and Self Assessment Sheet**" and upload it within iLearn. This is also a confidential document and will remain confidential between the lecturer(s) and each student. This will be used for identifying the contributions of each student to the project.
5. Finally, each student to upload their answers to **Data Mining Self Assessment Sheet** questions within iLearn. The answers to question 5 could be used to attract future students and/or make changes to the unit or for research on learning and teaching.
   * Points 3& 4 also apply to external students working in groups.

Poster Due is due on Week 13.

1. One poster per group (internal)/student (external) on iLearn by due date (power point document or pdf) clearly stating the group name (internal) or pseudo name (external).
   The posters will be available to all students so that they can be used for peer evaluation.

Note: The presentations of the project posters will take place in Week 13. Each student will mark the posters of other groups than their own group poster. Marking guide for the posters is provided in this document. The contribution of your fellow students' grading will be no more than 5% towards your unit mark. A similar study was undertaken by A/Prof Bilgin in 2006 semester 2 which is presented as a research paper in 2007. This paper can be downloaded from
http://www.stat.auckland.ac.nz/~iase/publications/sat07/Bilgin_Fraser.pdf

Modification to comply with the new Assessment Policy (Group work Students Only):

The new assessment policy states that "At least 50% of group work assessment shall be allocated to individual performance." https://staff.mq.edu.au/work/strategy-planning-and-governance/university-policies-and-procedures/policies/assessment-in-effect-from-session-2-2016

Therefore for students who will be working in groups (especially internal students), in addition to above specifications, one Appendix for each group member should be part of the submission. This appendix should be **no more than 3 pages per each group member** and should include the following:

Methodology, results and conclusion for the methods applied to data set, assuming each student in the group will apply at least one of the classification algorithms to the data set. If more than one applied by each group member, then choose the best within your own algorithms, if only one applied then write the results and conclusion as if this was the best algorithm.

These appendices will be marked as below:

Methodology	a) The suitability of the methods used for the analysis are explained b) The options chosen to create the classification models are specified (can be replicated by another researcher) c) The reasons for chosen options are provided	/5
	d) (if possible) Data mining evaluation is done by checking class specific error rates, lifts (category specific) and other means (i.e. gain, lift charts, sensitivity, specificity) e) (if possible) The best model is identified	/5
Results	a) The results for the best model is presented b) The results are discussed in the context of the problem identified in the introduction c) The results (i.e. decision tree) are explained correctly	/10
Conclusion	a) The most important findings are summarised b) How the findings could be used in day-to-day operations is suggested c) The limitations of the data set and/or analysis (if there were any) are discussed d) Future research is suggested	/5[1]
	Total Individual Mark	25

[1] If evaluation and identification of best model is not possible (e.g. student only applied one method) then conclusion could be reweighted to 10 marks.

Your Experience of working in a group for this Project

(For individual Submission for Internal Students) (Confidential)

1. Did you find easy or hard to work in a group? Why?

2. What problems did you encounter while completing this assignment as a group and how did you overcome them?

3. What kind of help would have been useful to you to improve the group dynamic?

4. In your view which participants made the best, most helpful or most useful contributions to the project? Why were these contributions so worthwhile?

5. In your view which participants made the worst, least helpful or least useful contributions to the project? Why were these contributions so irrelevant or unproductive?

6. Which part of the report did you contribute most?

7. Is there any part of the report that you did not contribute at all? If so, why?

8. How useful did you find the group work? (Rate from 1 to 5 where 1 is the least useful and 5 is the most useful)

9. What further comments do you have in relation to any of the questions above or any other aspects of the class?

10. If you are asked to grade your group's report, what would be the grade for this report? (see page 11 for the descriptions of the grades) Write one sentence to justify the chosen grade.

11. Please complete the Peer and Self-Assessment Sheet (next page) to help us to identify the individual contributions to this project. Do not forget to enter your own contribution (self assessment).

Peer & Self Assessment Sheet for Internal Students

Rate each member (including yourself) based on their contributions to each criteria.

			1	2	3	4	5	
			Little or no contribution	Below average contribution	Average contribution	Above average contribution	Outstanding contribution	
			Assessment Criteria					
			Logistics	**Leadership**	**Group Dynamic**	**Intellectual Contribution**	**Research, Writing, Editing**	
	Student ID	**Student Name**	Did s/he participate in group meetings in an organized fashion & meet group deadlines?	Did s/he provide leadership through listening to others and helping the group to function as a team?	Did the person share group responsibility without argument or disruption?	Did s/he provide useful ideas, helpful suggestions and feedback to the group?	Did s/he help in researching and writing the final paper in accordance with the group decisions?	**Totals for each student**
Yourself								
Group Member								
Group Member								
Group Member								
Group Member								
Group Member								

STAT828 Data Mining Self-Assessment Sheet (For individual Submission)

Please be as honest as you can be when answering the following questions:

1. In one sentence, what is the main point you are trying to convey in you report?
2. If you had additional time to work on this report/poster, would you want to change it? Explain.
3. What do you like most about your project?
4. What do you like least? Why?
5. You are asked by a friend to give advice regarding STAT828. Your friend wants to enrol the subject next year and not sure whether this is the subject that would be useful to her/him. What would you tell her/him?
6. Is there anything else that you would like to mention?

This assignment is designed to help you develop the following graduate capabilities:

1. Discipline knowledge and skills (identify and apply appropriate data-mining techniques to a new problem; demonstrate the use of the freeware package “R” and/or IBM SPSS Modeler in carrying out some of these data-mining techniques)

2. Critical, analytical, integrative and creative thinking (identify and embed a graphical display when it is appropriate; identify the suitable data pre-processing methods for a new problem; and apply a suitable data mining technique to current problem; examine and compare the differences between different models and interpret the results from sophisticated models to end users. In addition, discuss the limitations of such a method)
3. Research and problem solving capabilities (identify which variables will be cleaned, transformed and apply suitable methods to deal with these and similar problems (i.e. outliers, missing data); identify a business problem for a data set and propose a solution for the problem by applying suitable data mining methods)

4. Effective communication (written communication using formal language and academic conventions, particularly report writing skills)

5. Socially and environmentally active and responsible (prepare a professional report for the results of the analysis where an action plan for the management is outlined)

6. Professional and personal judgment and initiative (analysis, discussion and conclusions; and self-assessment)
7. Engaged and responsible, active and ethical citizens (when making decisions to include variables for analysis, consider the implications for the local or the global society; report any outliers that were excluded from the study and explain why)

   Commitment to continuous learning (identify R packages or similar free software that could be used for analysis and then learn them yourself to be effectively solving the problem in hand; share the learning experience with peers in class)

The Criteria for the Project Report

HD: There is substantial originality and insight in identifying, generating and communicating competing arguments, perspectives or problem solving approaches; critical evaluation of problems, their solutions and their implications; creativity in application as appropriate to the discipline.

D: There is demonstration of frequent originality in defining and analysing issues or problems and providing solutions; and the use of means of communication appropriate to the discipline and the audience.

Cr: There is demonstration of substantial understanding of fundamental concepts in the field of study and the ability to apply these concepts in a variety of contexts; convincing argumentation with appropriate coherent justification; communication of ideas fluently and clearly in terms of the conventions of the discipline.

P: There is demonstration of understanding and application of fundamental concepts of the field of study; routine argumentation with acceptable justification; communication of information and ideas adequately in terms of the conventions of the discipline. The learning attainment is considered satisfactory or adequate or competent or capable in relation to the specified outcomes.

F: There is missing or partial or superficial or faulty understanding and application of the fundamental concepts in the field of study; missing, undeveloped, inappropriate or confusing argumentation; incomplete, confusing or lacking communication of ideas in ways that give little attention to the conventions of the discipline.

The Marks for each part of the Project Report

Executive Summary:	/5
Introduction:	/5
Description of the data set:	/10
Methodology:	/10
Results:	/10
Conclusion:	/5
Other:	/5

Note: The marking scheme is provided as part of this document (last two pages).

Poster Presentation Marking Guide

Poster Title:____

Name of group being marked:______________________________

Assessor’s Name: ______________________________

Assessor's Group: ______________________________

Clarity of the message

The poster is characterized by three important elements:

1. Language is clear and understandable
2. Stated objective(s), research question(s) or hypotheses
3. Structure of the presentation is clearly and logically set out.

Not Present	Low	Mid	High
There is no clear message. None of the three elements are present.	Only one of the three elements above is present. OR All three elements are of low quality.	Only two of the three elements above are present. OR All three elements are present but they are of mid quality.	All three of the elements are present and they are of high quality.

Understanding of the Data Set

The poster is characterized by three important elements:

1. Variables in the data set are identified and described, data source identified
2. The set data is appropriate for answering the research question
3. Quality of data is addressed (i.e. accuracy of measurements, data preprocessing, missing values, outliers)

Not Present	Low	Mid	High
None of the three elements are present.	Only one of the three elements above is present.	Only two of the three elements above are present.	All three of the elements above are present.

Analysis and conclusions

The poster is characterized by three important elements:

1. Analysis is appropriate to answer the research question and for the kind of data
2. Conclusions are stated and supported by the results
3. Limitations are discussed or improvements are suggested

Not Present	Low	Mid	High
None of the three elements are present.	Only one of the three elements above is present.	Only two of the three elements above are present.	All three of the elements above are present.

Graphs and tables

The poster is characterized by three important elements:

1. The type of graphs/tables is appropriate for displaying and summarizing the data.
2. The number of graphs/tables is appropriate (adequately show different perspectives and the information is not redundant).
3. Graphs/tables are properly titled and explained.

Not Present	Low	Mid	High
None of the three elements are present.	Only one of the three elements above is present.	Only two of the three elements above are present.	All three of the elements above are present.

Presentation The poster is characterized by three important elements: 1. It can be read from a reasonable distance (i.e. font is not too small) 2. There is a good balance between graphs and text 3. Appearance is neat and eye-catching			
Not Present	**Low**	**Mid**	**High**
None of the three elements are present.	Only one of the three elements above is present. OR All three elements are of low quality.	Only two of the three elements above are present. OR All three elements are present but they are of mid quality.	All three of the elements are present and they are of high quality.

Creativity/Importance The poster is characterized by three important elements: 1. The research question is creative or original 2. The study answers an interesting question 3. The design is creative and original			
Not Present	**Low**	**Mid**	**High**
None of the three elements are present.	Only one of the three elements above is present.	Only two of the three elements above are present.	All three of the elements above are present.

Which of the following grades do you feel BEST reflects the quality of the poster? (Circle one)

HD	**D**	**Cr**	**Pass**	**Fail**

Any other comments:

The Marks for each part of the Project Report **Group Name:__________________**

Part of Report	Details covered	Mark
Executive Summary	a) Sets the scene (background) b) Specifies what has been done (methodology) c) Summarises results or presents most important results d) Informative	/5
Introduction	a) Background to project is described (not general terms) b) Problem identified c) The aims of the project introduced d) Intended users or use of the solution is identified e) Following sections listed with what is in them (overall structure of the report)	/5
Description of the data set	a) Errors are identified and if possible, corrected in the data set b) Outliers are identified, if any, and a solution suggested c) Missing data is identified, if any, and a solution suggested d) If new variables are created, the reason(s) and how they were created are explained e) If concept hierarchies are created the reason for creation is explained f) If any observation is excluded from analysis, the reasons are provided g) If any variable is excluded from analysis, the reasons are provided h) Descriptive data mining is performed (i.e. summary statistics for variables and/or graphical displays) and a summary is written i) The relationships between the outcome and predictor categories are explored	/10
Methodology	a) The suitability of the methods used for the analysis are explained b) The options chosen to create the classification models are specified (can be replicated by another researcher) c) The reasons for chosen options are provided d) Data mining evaluation is done by checking class specific error rates, lifts (category specific) and other means (i.e. gain, lift charts) e) The best model is identified	/10
Results	a) The results for the best model is presented b) The results are discussed in the context of the problem identified in the introduction c) The results (i.e. decision tree) are explained correctly	/10
Conclusion	a) The most important findings are summarised b) How the findings could be used in day-to-day operations is suggested c) The limitations of the data set and/or analysis (if there were any) are discussed d) Future research is suggested	/5
Other	a) Pages are numbered b) Report has no more than 10 pages (excluding appendices) c) Individual parts are included as appendices (no more than 3 pages per student). d) All tables and graphics have informative captions e) All tables and graphics are explained and referred in the text f) No (minor/major) spelling/grammatical mistakes g) At least three references are listed in the reference list h) Computer outputs and R codes are used sparingly in the body of the report i) R script and/or IBM SPSS Modeler stream and/or IBM SPSS Statistics output files (.spv) are submitted j) Experience working in a group is submitted k) Peer and Self Evaluation is submitted	/5
	Total Mark	/50

Resources to might be useful

Developing a Poster Presentation

http://www.kumc.edu/SAH/OTEd/jradel/Poster_Presentations/PstrStart.html

Microsoft Office Site (search for poster)

http://office.microsoft.com/en-au/templates/default.aspx

Scientific Poster Design from Berkeley

http://hsp.berkeley.edu/sites/default/files/ScientificPosters.pdf

StudyWISE is an iLearn resource created by Learning Skills at MQ

http://students.mq.edu.au/support/learning_skills/studywise

(Let me know if any of the links are broken)

IBM SPSS Modeler 18.1 student access (discovered by a student in 2018)

If you'd like your own copy (for Windows - I haven't explored Mac options):

- Go to https://estore.onthehub.com
- Create an account using your @students.mq.edu.au email address.
- Once your account is active and logged in, search for "Modeler"
- Add "IBM SPSS Modeler Premium Academic and Faculty/Author 18.1 Microsoft Windows Multilingual eAssembly - Students" and "SPSS Modeler v18.1 - Student License key" to your cart. They are both free.
- Checkout.
- You'll need to answer some questions about the course you are enrolled in.
- You'll be given a list of download options. You'll need to download a Modeler 18.1 client (not the "Premium" version, as it won't work until you have the basic version installed). I downloaded "IBM SPSS Modeler Client 64-bit 18.1 Microsoft Windows Multilingual (CNKI1ML)".
- Unzip the file you have just downloaded, and run Setup.exe.
- Once installation is complete, run IBM SPSS Modeler 18.1. You'll be told your license has expired.
- Input the activation key you have been provided when you "purchased" the free student license key.
- You now have a one-year, free license to Modeler 18.1