Group Project Guideline

1 Objective

The group project is intended to: 1) give you an opportunity to work on a real-world dataset; 2) let you apply the theoretical and practical concepts covered in class (and beyond via additional readings or independent research); 2) promote collaboration between team members to mimic business settings.

2 Summary of Deliverables

The project component accounts for 20% of the final grade. Further breakdown is listed below.

Deliverable	Weight	Deadline	Description
Team formation	-	10 Feb 2020	3-4 members
Proposal	5%	4 Mar 2020	max. 1 page, min. font size 10
Final report	5%	8 Apr 2020	max. 5 pages, min. font size 10
Codes	10%	8 Apr 2020	Reproducible, modular, and well documented Python Jupyter Notebooks

3 Project Content

You project should generally follow the main steps below.

- Find a dataset from the UCI ML repository (see more details under “Datasets” below);
- Formulate an objective or motivation for your project (What do you want to investigate?)
- Perform one or more analytical tasks (see more details under “Potential analytical tasks” listed below);
- Present results using appropriate visualizations (e.g. tables, plots, charts, etc.);
- Summarize the significance of your results (e.g. What are the insights learnt? Is your objective achievable, or is your motivation valid? What is your conclusion? Are they intuitive, or surprising, and why?).

Datasets

You should work **only on one dataset** from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php), but you may perform additional filtering to only focus on a smaller sub-dataset of your interest. When choosing your dataset, please use the below criteria

1) Data Type: must be “Multivariate”
2) Default Task: must contain either keyword “Regression” or “Classification”
3) Attribute Types: for regression, the response variable must be numerical; for classification, the response variable must be categorical. Depending on your objective or motivation, use your own judgment to choose the predictor and response variables, possibly after some exploratory data analysis.
4) # instances: between 100 and 10,000
5) # attributes: between 10 and 100

Potential analytical tasks

Below are the suggested tasks you could perform on your dataset. It is important to note that it is not enough to merely perform the tasks; you need to investigate and interpret your results to derive insights. For example, is your objective or motivation validated? What conclusions can you reach based on your results? Are the findings intuitive? Is there any surprising finding?

You are NOT required to perform all of them—as few as one is enough, if it can clearly support your insights or conclusion. Of course you may explore multiple tasks in order to strengthen your conclusion, in which case those multiple tasks should be related to support each other rather than being independent tasks. For instance, after performing an extensive EDA, it may become evident that some predictors have a linear relationship with a given response variable. You might then want to build a linear regression model.

The potential analytic tasks are as follows:

- *EDA*: find an interesting insight or trend via extensive EDA including visualization;
- *Regression*: formulate an appropriate and interesting problem statement involving predicting the numerical value of the response variable;
- *Classification*: formulate an appropriate and interesting problem statement to predict the class type of the response variable;
- *Others:* Hypothesis testing, Bayesian modelling.

Note that note some of the tasks are only going to be covered after the proposal is due. It is possible for you to add on additional tasks when it is appropriate in your project even though you might not have included it in your proposal.

4 Proposal

Your proposal must not exceed 1 page with minimum font size 10, comprising of the following:

1) Project title, group name, and group member names.
2) Objective or motivation: What do you intend to investigate? Include any background necessary to understand your goal.
3) Dataset: Identify and describe the dataset in your proposal.
4) Approach: Describe the potential tasks to perform, and how do you plan to achieve your objective and/or validate your motivation?

5) Timeline: List a schedule to achieve major milestones, breaking down into weeks.

Your proposal will be graded based on its **interestingness**, **feasibility** and **clarity**. Feedback will be sent to you via eLearn. If there is any major issue, please discuss with the instructors.

5 Final Report and Codes

Your final report must not exceed 5 pages with minimum font size 10, comprising of the following:

1) Cover page: project title, group name, and group member names (not included in the 5-page limit);
2) Abstract: A short summary of your project in around 100—200 words.
3) Introduction: Introduce your objective or motivation—What do you intend to investigate? Include any background necessary to understand your goal.
4) Dataset: Describe the dataset. What are the data about? What filtering, sampling or any other transformation/pre-processing steps have you performed on the data?
5) Approach: Describe the analytical task(s) you performed, including the input and the output, the algorithmic steps, and any configurations or parameter settings, as well as the tool(s) used.
6) Results and discussion: Describe the results you obtained. Include further discussions, such as the insights you have drawn from your results. What is the significance of your results? Are they intuitive or surprising and why? Any potential future study based on your current results?
7) Conclusion: Did you achieve you goal? What have you done right or wrong? Typically short in around 100—200 words.
8) Contributions: Describe the contribution of each member, e.g. who did what, including both tangible (e.g. implementation and report writing) and intangible (e.g. generating ideas, planning).
9) References: list any references or sources you have cited

Throughout the report, especially in the sections of dataset, approach, result and discussion, remember to include appropriate visualizations (e.g., plots, charts and tables) to illustrate your idea. Your final report will be graded based on its **interestingness**, **validity and soundness**, and **clarity**.

You final report should be accompanied by your codes. Please submit a zipped file containing the report, Jupyter Notebook and the dataset you used, and ensure the following:

1) Write modular codes.
2) Document your codes and provide comments and/or markdown cells to facilitate easier understanding of your codes.
3) Results in your reports must be reproducible in your notebooks.

7 Peer Evaluation

At the end of the term, a peer evaluation will be conducted via eLearn. In the event of any unusual peer evaluation (i.e. a member receiving very low score with detailed comments supplied), the instructor reserves the right to conduct further investigation and moderate the score of each team member accordingly based on the investigation, if necessary. Your evaluation will be anonymous and confidential to all other team members or students, although the teaching team will be able to see your identity in order to conduct any investigation and/or moderation. More details will be released towards the end of the term.