Data Mining – Project (mandatory for grad students, optional for undergrad students)
(Forming a group is recommended. Groups of up to 3 are allowed.)

This is an exploratory project. You are encouraged to collect or find interesting data sets for an application domain that interests you. The more data you have the better it is for finding interesting patterns. Your project should consist of the following three phases:

Data Collection (for the domain you like/are interested in)
Data Preprocessing and Visualization (data cleaning and transformation into a useful form)
Data Mining (using algorithms you have seen so far)

At least one of these phases should be not trivial. For example, the data collection and preprocessing phase could be non-trivial (e.g. the sites you use have some specific APIs that you need to use and/or the data needs special preprocessing).
Or the data mining process could be non-trivial. E.g. you collect a lot of data and then some algorithms such as those in scikit-learn, being main memory algorithms, have a hard time to mine your data. In such a case, you should modify or recode those algorithms, or research available algorithms that can be more efficient for large amounts of data.

You should submit a report describing your work. The length of the report should approximately be 10 pages. The report and your work can also be all in a jupyter notebook, which you submit along with your data.

Depending on the number of groups, there may not be enough time for every group to present in class. If this happens, only a selected subset of groups will be invited to present their projects during the last weeks of the term. However, all groups are required to prepare a presentation (without a voiceover) to accompany their project report.

Some interesting data/articles references are:

http://www.kaggle.com/c/titanic-gettingStarted (but do not choose this is for a project)
http://www.kaggle.com (you can choose a non-trivial project there)
https://www.kdnuggets.com/2016/11/rank-ten-precent-first-kaggle-competition.html
https://towardsdatascience.com/how-i-ranked-in-the-top-25-on-my-first-kaggle-competition-9ea53499d58d
http://www.sciencedirect.com/science/article/pii/S1877050916309036
http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD
http://labrosa.ee.columbia.edu/millionsong/pages/tasks-demos
http://archive.ics.uci.edu/ml/datasets/URL+Reputation
http://cseweb.ucsd.edu/~voelker/pubs/mal-url-icml09.pdf
https://www.yelp.com/dataset
https://grouplens.org/datasets/movielens
https://conf.researchr.org/track/msr-2023/msr-2023-mining-challenge

Project Rubric

	Weight	0-50%	50-75%	75-100%
Questions	**10**	Questions overly simplistic, unrelated, or unmotivated	Questions appropriate, coherent, and motivated	Questions well motivated, interesting, insightful, and novel
Analysis	**60**	Choice of analysis (exploratory data analysis, feature engineering, model selection, training, tuning, metrics, etc.) is overly simplistic or incomplete. Inappropriate choice of plots; poorly labeled plots; plots missing.	Analysis appropriate. Plots convey information but lack context for interpretation.	Analysis appropriate, complete, advanced, and informative. Plots convey information correctly with adequate and appropriate reference information.
Conclusions	**10**	Conclusions are missing, incorrect, or not based on analysis.	Conclusions relevant, but partially correct or partially complete.	Relevant conclusions explicitly tied to analysis and to context.
Writing	**10**	Explanation is illogical, incorrect, or incoherent	Explanation is correct, complete, and convincing	Explanation is correct, complete, convincing, and elegant
Presentation	**10**	Presentation is illogical, incorrect, or incoherent.	Presentation is readable and clear.	Presentation is appealing, informative, and crisp.
	100			