Data Mining – Project
(Forming a group is recommended. Groups of up to 3 are allowed.)

This is an exploratory project. You are encouraged to collect interesting data sets for an application domain that interests you. The more data you collect 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 Weka, 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; this is a preferred way to do the project because everything is transparent, and we can re-run your work. However, you don't have to. You can use any language and tool you prefer.

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://dnc1994.com/2016/05/rank-10-percent-in-first-kaggle-competition-en
http://www.sciencedirect.com/science/article/pii/S1877050916309036
http://archive.ics.uci.edu/ml/datasets/Farm+Ads
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
http://www.yelp.ca/academic_dataset
https://grouplens.org/datasets/movielens/
http://2013.msrconf.org/challenge.php
http://2015.msrconf.org/challenge.php
http://2017.msrconf.org/#/challenge
https://2018.msrconf.org/track/msr-2018-Mining-Challenge