CS249 – Principles of Data Mining/Data Science – Spring 2014
Course Project

A course project is basically one or more IPython notebooks implementing a log of analyses of your data. The results can include visualization output, descriptive statistics, algorithms or models developed, etc.
The project also includes some documentation — a summary saying who did what and giving lessons learned, a set of final presentation slides, and a github site with the project files.

Some examples of past projects: analysis of strategies in baseball (also soccer, etc.), what kinds of movies win the Golden Globe awards, pairs trading stock market analysis, how countries try to win gold medals in the Olympics, which kinds of cars get EPA approval, seasonality of global tourism, rock music classification.

Work on data that is interesting and that you care about. The more interesting the data, the more interesting the story or results that you can find in it. Once you have data you care about, it is fun to put work into analyzing it, and it is easier to find patterns.
If you are not sure where to find interesting data, but know people that you are interested in working with (and who have data — e.g., UCLA faculty, or friends at Google or Yahoo, a company you want to do an internship with, etc.), you could try starting with that.

The data should be at least in the multi-megabyte range. No toy datasets. However, it is often a bad idea to jump into a multi-gigabyte dataset immediately — instead it may work better to start on samples, and analyze the full dataset when the notebook is working.

Groups of up to 5 students are encouraged to work together on a challenging project.

Project Deliverables:

The project should produce a set of specific things:

- a github site — that is, a page at `http://github.com`. You can sign up for an account at `https://github.com/join`. (What exactly is github anyway?) Project files should be stored on this site, so that others can re-run the notebook. You can use any programming languages you like — R, Python, whatever.

  Github hosts many R projects; e.g. Hadley Wickham's Advanced R development site mentions github best practices.

- one or more IPython Notebooks that include:
  - a description of a dataset you used, and how it was obtained;
  - a step-by-step log of analyses you attempted; There is no way to guarantee data mining will find results, but the log can record your exploration.
- a PDF *presentation* describing the project — e.g., a set of PowerPoint slides, or a set of beamer LaTeX slides.
- A PDF project summary document that lists:
  - a summary of experiences, insights, and lessons learned.
  - a summary of who did what for each of the grading criteria below.

Grading:

The project will be graded mainly on effort invested. Think of it as something to put on your resume.
Grading criteria:

- *Data*: data acquisition effort, data novelty.
- *Analysis*: IPython notebook construction effort, data analysis effort, data analysis methods.
- *Documentation*: presentation effort, project summary effort, github site effort.
- *Overall*: project methods, project creativity, project difficulty.

The project summary should discuss all of these criteria, one-by-one. For Team Projects, it should say who did what.

Sample Student Projects using IPython Notebooks:

- Summaries of Student Project Notebooks from a Berkeley Data Science course.
- A video of presentations of Student Project Notebooks from a Harvard Data Science course.

Interesting Examples of IPython Notebooks:

- An inventory of many interesting applications of IPython – from online courses to reproducible science to executable books.
- The Wakari Notebook Gallery — with tutorials about Python-related data science tools — by http://www.continuum.io, the people who make the Anaconda distribution.
- The IPython.org gallery of interesting IPython Notebooks, including a collection of notebooks that can be browsed via nbviewer.
- Notebooks related to social network analysis, e.g.:
  - Fernando Perez' analysis of graph properties of the Twitter stream.
  - Mining Social Web APIs with IPython Notebook — online resources for an O'Reilly book about analyzing sources like Twitter: called Mining the Social Web. There is also a related online webcast video by the author, Matthew Russell, titled *Data Science Experiments with Twitter and IPython Notebook*.
- Notebooks and Datasets from *Python for Data Analysis*, our course text. Also, sample notebooks at Wes McKinney's site, including a notebook analyzing Stackoverflow posts about Python (with a dataset of posts), and a tutorial notebook about Pandas (with an associated downloadable data zip file).
- Google Search using **ipynb** as a search term can also find all kinds of notebooks. e.g., ipynb linkedin, ipynb yelp, etc.

Sample Data Sources:

- Google Public Data Server http://www.google.com/publicdata/directory
- The list of Datasets http://www.kdnuggets.com/datasets/ at kdnuggets.com — a hub for data mining.
- InfoChimps — a good starting point for finding data: http://infochimps.org
- Kaggle — interesting data and data mining contests: http://www.kaggle.com
- Amazon EC2 Public Data Sets (huge datasets for genomics, Wikipedia, economics, astronomy, etc.): http://aws.amazon.com/datasets
- An amazing epidemic: http://www.cdc.gov/obesity/data/adult.html See also the Google Correlate history, showing the history of search activity
- Baby Names dataset (Social Security card applicants since 1879): http://www.ssa.gov/oact/babynames/background.html
- The Million Song Dataset: http://labrosa.ee.columbia.edu/millionsong/
  e.g.: http://labrosa.ee.columbia.edu/millionsong/pages/matlab-introduction
- Datasets described in the Doing Data Science book.
- Guardian.co.uk/data http://www.guardian.co.uk/data
- NY Times Labs http://nytlabs.com
- Los Angeles Times Data Desk http://datadesk.latimes.com
- Federal data clearinghouse http://www.data.gov
- National Bureau of Economic Research http://www.nber.org/data
  (many interesting datasets: Macroeconomics, industry, trade, demographics, hospital, patents, ...)
- Federal Reserve Data Economic Research & Data http://www.federalreserve.gov/econresdata/default.htm
  (including data about mortgage defaults, interest rates, exchange rates, industrial production, ...)
- Federal Statistics Data Access Tools: http://www.fedstats.gov/toolkit.html (see also: http://data.gov)
- The Federal Election Commission's 2012 Campaign Contribution data, a 150 megabyte CSV file that includes contributor names, occupation and employer, address, and contribution amount. (wakari.io has a IPython Notebook that links this FEC data with Census data to obtain per-capita contributions by state.)
- Tracking the U.S. Congress http://www.govtrack.us/developers/data.xpd
- California State Datasets http://www.ca.gov/data/state_data_files.html