Group Project

Group Project (40%): Students will form groups of 4-5 and jointly analyze a dataset "in the wild." Students may either form groups of their own choosing, or may ask the instructor to be randomly paired with other group members.

The goal of the project is to take an original dataset and analyze it, bringing to bear all the might of your newly acquired analytic muscle. Students should seek to perform a novel analysis on their chosen dataset, to reveal previously undiscovered patterns and answer previously unknown questions. Examples of datasets which students may choose to analyze include, but are not limited to:

There are also several extensive lists of datasets that you might look to for other ideas:

- Twitter data (cleaned data available from http://twitter.mpi-sws.org/, http://blog.infochimps.org/2008/12/29/massive-scrape-of-twitters-friend-graph/)
- Wikipedia (http://wiki.dbpedia.org)
- Movies database (http://www.linkedmdb.org/)
- Government data (e.g. https://watchdog.jottit.com/volunteer?r=83)
- Yahoo data (http://webscope.sandbox.yahoo.com/index.php)
- Mobile phone data (http://www.d4d.orange.com/home)
- Best Buy (https://bbyopen.com/developer)
- Medical data (http://blog.wolframalpha.com/2010/06/29/disease-and-patient-level-statistics-with-wolframalpha/)
- Flickr (http://www.isi.edu/~lerman/downloads/flickr/flickr_taxonomies.html)

The questions which you seek to answer in working with your dataset are up to you. However, your project will be evaluated, and your grade determined, using the following criteria:

Dataset identified: 10 points (DUE July 3)

Question identified: 10 points (DUE July 7)

Preliminary Analysis: 20 points (DUE July 10)

Final Presentation: 30 points (DUE July 15)

Final Paper: 80 points (DUE July 17)

- Late submissions for any of the above will be accepted only 1 day late with a 20% deduction. After 1 day, the assignment will not be accepted.

This project has the following deliverables:

- **<u>Dataset</u>** (1 paragraph), due by class on **July 3:** In one paragraph, explain what dataset you are going to explore and why.

- **Project Proposal** (1 page), due by class on **July 7**: Perhaps the most difficult part of this assignment is properly scoping your analysis. In 1-page explain the question you intend to investigate with your data. Explain how you intend to explore it. Note any limitations or problems you may run into with the data.

- **Preliminary analysis and summary statistics** (several tables and graphs), due by class on **July 10**: Conduct a first-pass descriptive analysis of your dataset. Produce tables and graphs that show exactly what data you have, and that contain summary statistics about the data. Questions to answer for each data source include:
  - How many unique observations to you have?
  - What information/features/characteristics do you have for each observation?
  - What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)?
  - Are there obvious trends in the data (over time, across subgroups, etc.), and are the differences statistically significant?
  - What are the other salient aspects of the data (e.g. geospatial factors, text content, etc.)
  - Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.

- **Final report** (4-5 pages), due in class on **July 15**: The final report will be the primary basis for evaluation of your project. The first page of your report should consist of a title page that lists the names of all group members, a title for your project, and a short abstract, no more than 250 words, which describes the key findings of your group. Include in your report any figures, tables, or multimedia (as hyperlinks) that you develop in the course of your project. Code and other relevant but non-essential material may optionally be included as appendices. The content of the title page, references, footnotes, and appendices does not count toward the word limit. In the body of your report, you should include a discussion of the following:

  - Introduction/Background of Project and Research Question
  - Methods (including details about the data)
  - Results
  - Discussion
  - Conclusions and Future Directions
  - References

  Your Final paper will be evaluated according to the following criteria.

  1. Timely submission (10 points)
  2. Overall ambition, creativity, novelty, and difficulty of project (20 points)
  3. Clarity of empirical questions (10 points)
  4. Analysis (10 points)

5. Use of appropriate and compelling visualizations (10 points)
6. Clarity of writing and organization of written report (20 points)

- **<u>Final Presentation</u>** (~ 8 minutes) in class on **July 17**. The final presentation is your chance to show the class what you have been working on for the last three weeks. I encourage questions from the audience during the presentation. Your presentation will be evaluated according to the following criteria.

  Your Final presentation will be evaluated according to the following criteria.

  1. Motivation and Background of Project (5 points)
  2. Research Question (5 points)
  3. Methods and Data used (5 points)
  4. Results and Interpretation (10 points)
  5. Conclusion and Future Work (5 points)