CS1070: Taming Big Data

Final Group Project: Exploratory Data Analysis

A. PROPOSAL DUE: **Tuesday, November 26th BEGINNING OF CLASS**
B. FINAL PROJECT DUE: **Thursday, December 12, 12:00 – 1:50PM**

Task:

In groups of 3-4, perform an original, exploratory data analysis on a data set of your choosing. The data set can come from any source as long as it is something you have not already worked on for this course. Possible sources of data include but are not limited to:

- Open data sources such as Data.gov, OpenDataPhilly.org, and OpenData by Socrata.
- Data sets from the Pew Research Center (www.pewresearch.org/data)
- Sports statistics, such as those from the Baseball Archive (http://www.seanlahman.com/baseball-archive/statistics)
- An original survey conducted by your group.

Your analysis should clearly demonstrate the tools and techniques you've been exposed to in this course. This can take any form you'd like but should answer some interesting question(s) about the dataset. Your analysis must include at least one of the following topics from the later part of the semester:

- Sample Comparison & Hypothesis Testing
- Prediction
- Classification

You must additionally make use of data visualization tools learned throughout the semester to drive your analysis.

Deliverables:

A. Proposal

As a group, you will submit a 100-word proposal by the due date (please refer to the respective due date of the proposal at the top of this document).

This proposal should contain answers to what, why, and how questions. Specifically, you will briefly explain what you are studying, why you are studying it (why it is important or why we should care about it-- mainly the motivation behind the problem), and how you will be studying it (your data source, data tool you will use, and methodology you will follow).

The purpose of this proposal is for me to have a chance to assess the value and feasibility of your project. Note that at this stage I do not expect you to have broad and/or in-depth knowledge on the chosen topic, therefore it is perfectly OK if you cast doubts or raise questions in the proposal. However, I may ask you to change your chosen topic if it sounds undoable.

Do not exceed 100-word in the body of the proposal (this limit excludes titles or team member names). One submission per group will be sufficient.

B. Final Project

(i) A scientific poster
(ii) The data you used (or pointers to it)
(iii) Your Jupyter notebook containing ***well commented*** code
(iv) In class presentation of (i)

Your poster must include the following information

- Your group members, and the title of your presentation.
- What question do you want to answer and why is it important?
- What data are you analyzing? What are the key elements and how did you get it?
- Your analysis and the results. Make good use of data visualizations.
- Your conclusions. What did you learn? Support your conclusions using the results of the analysis, citing specific evidence!
- A list of any references you used.

Your group will present your work a poster session during the allotted presentation time. At any time, at least one group member will stand by the poster and explain it to anyone who walks up; group members should rotate who is presenting, and go visit other posters when they are not presenting their own work.

Final Project - Grading:

There are three steps to grading your individual scores for this group project.

First, you will receive 10% for your project proposal being submitted, and completed on time. Each day late will lose you 2.5%.

Second, as a group, your project will get a single score from your instructor using the evaluation criteria attached on the next page. This score will be the same for all group members and will account for 50% of your final project score.

Second, you will submit a peer review of your group members online -- the submission link will become available when presentations are over. Submitting peer review will account for 10% of your final project score. If you don't submit you will get 0 from this step.

Third, average score you get from your group members' peer reviews will account for 30% of your final project score.

Overall, the score your group project gets from your instructor, submitting a peer review, and average peer evaluation you get from your group members will determine your final project score. For instance:

1. First you will receive 10% simply for completing your proposal on time.
2. Using the criteria on next page, your instructor determines that you get 45 out of 50.
3. You submit the peer review of your group members and you get 10 points.
4. Your peers give you an average score of 25 out of 30.
5. Your final (individualized) project score then becomes 90 out of 100 (=10+45+10+25).

EVALUATION CRITERIA FOR FINAL PROJECT

TRAIT	4 (A-level)	3 (B-level)	2 (C-level)	1 (D or F-level)
Scenario Identification (20%)	Comprehensively and clearly identifies and describes the issue. The central question is clearly stated.	Precisely identifies and describes the issue including the majority of its key components/variables. The central question is somewhat clearly stated.	Correctly identifies issue but certain key components/variables remain unclear or omitted. The central question is somewhat unclear.	Limited ability to clearly identify the issue and its various components/variables. The central question is unclear or missing.
Data Gathering (20%)	Gathers correct and highly credible data that directly answers the central question.	Gathers correct and highly credible data that mostly answers the central question.	Gathers relevant data that somewhat answers the central question.	Gathers and incorrect, insufficient or unreliable data. Data is not directly related to central question.
Analysis (20%)	Presents an insightful and thorough analysis. Analysis is driven by logical arguments clearly related to central question.	Presents an effective analysis. Analysis is driven by arguments that, while slightly flawed, are related to central question.	Presents a superficial analysis of central question. Aspects of the question are unanswered.	Presents an incomplete analysis of central question. Analysis is not related to the central question.
Conclusions (20%)	Clearly identifies and articulates all implications and consequences of analysis. The conclusions are clearly related to the results of the analysis.	Somewhat clearly identifies and articulates all implications and consequences of analysis. The conclusions are mostly related to the results of the analysis.	Some implications and consequences of analysis are unidentified, unaddressed, or unarticulated. The conclusions are weakly related to the results of the analysis.	Implications and consequences of analysis are mostly unaddressed. It is unclear how conclusions follow from the analysis and what was learned from the analysis.
Visual Appeal (20%)	Visuals display high levels of creativity and strongly enhance the effectiveness of the presentation. Clearly leverages principles of good visualizations.	Visuals display satisfactory levels of creativity and generally enhance the effectiveness of the presentation. Mostly leverages principles of good visualizations.	Visuals display marginal levels of creativity and somewhat enhance the effectiveness of the presentation. Principles of good visualizations are underutilized.	Visuals do not enhance, or get in the way of, the effectiveness of the presentation. Principles of good visualizations are not used.