Instructor: David Brady
DS 52XO: Term Project

DS52X0 PROJECT GUIDELINES

DUE DATES AND LINKS (TENTATIVE)

- **Step 1:** Deadline to submit teams
    - Choose a team of max 2 students in your class.
    - Each student can belong to at most one team.
    - Single-member teams must be approved by a TA
    - Each team member must submit a complete list of team members
- **Step 2:** Discuss project ideas with TAs and the instructor, after you review the following sources of data
    - Must be publicly available data
    - Consider data science competition sites: Kaggle, DrivenData, CrowdANALYTIX, TunedIT, InnoCentive, Codalab
- **Step 3:** Abstract is due for grading (see later section for abstract guidelines)
- **Step 4:** Milestone report is due
    - Details will be provided soon
- **Step 5:** Project Report and Presentation due
    - Guidelines are presented below. More details will be provided soon.

BREAK DOWN OF ACTION ITEMS

ABSTRACT

The project abstract should not exceed one page in length with 11-point font and single spacing. The first half of the document will describe the publicly available dataset you intend to explore - describe the measurements carefully, so that the readers will see the connection between the goals for exploratory analysis and the available measurements. Present this section in your own words, without copying from archived manuscripts. Assume that the reader provides funding (I.e., with little technical background, but with funds that you hope to receive). Hence, the goal is to convince prospective funder that the project is feasible. For this, demonstrate that your team understands the available data carefully enough to pose interesting hypotheses and explore with at least some low-risk options and a few high-risk, high-reward teasers.

The final half of the document will describe specific goals for your exploratory data analysis (EDA). Write this from the perspective of a team leader, with limited resources and people to perform EDA. How will you gain knowledge from the data? What do you hope to do, specifically, and why? Like what is the application of an answer to that question, i.e., *so what, who cares?* What do you believe can be answered using methods developed in the class? For each question, present three procedures from class that you will apply to the data set. You should expect two-thirds of your procedures to fail (excellently) and half of your questions to be not be answered accurately enough.

Instructor: David Brady
DS 52XO: Term Project

Generate enough questions and procedures so that some procedures will answer some of your questions. Take risks but have a low-risk solution. You will have to document why procedures and questions failed at the Milestone report , so don't propose more than your team can address in the allotted time – the importance of properly planning (I.e., time management) is especially important in data science. The ability and know-hows are important, but as too is the project planning. Do not overlook this and fall into a trap of putting too much on the plate, or, to the contrary, not planning for work to match the potential of your team (I.e., undershoot expectation).

PROJECT REPORT

The project report will be your opportunity to tell the story from slides in a single, technical write-up. It should consist of the edited slide set, along with detailed notes to accompany each slide. Think of the project report as the script for the project presentation (next). The report format will allow one 8x11 page for each slide: the slide itself will occupy 1/3 of the page, and the notes will complete the page (and no more). When reading the notes aloud at a presentation pace, the report page should take 1.5 minutes on average. Each member of the group will contribute an equal portion to the report.

PROJECT PRESENTATION

The work of each team will be presented as a recorded talk - lasting up to 10 minutes and recorded by the team. Each team member should spend about the same amount of time speaking during the presentation. When recording the talk, imagine your audience: recruiters, peers, TAs, and faculty. Rehearse your talk and speak to this audience. Described what methods were more successful and devote more time to these strategies. Mention which techniques were less successful, and why you believe they were so.

Your presentation will be graded by me, as well as multiple reviewers from DS5220 and DS5230 (next). A grading sheet will be distributed to establish consistent criteria for grading, and outlier scores will be removed. The grading sheet will be provided later.