DS-GA 1007
Programming for Data Science

Professor Jeremy Curuksu
Center for Data Science
New York University
Email: jeremy.cur@nyu.edu

Project

Students are asked to work in group of 2 (minimum) to 4 (maximum) and propose a problem to work on. Examples of proposal will be presented early October. Projects will be graded by a panel of instructors who will assess several dimensions (listed below), based on both the code and a presentation in person.

Grading:

The project will be graded based on:

- **20**%: Demonstrated understanding of Python concepts learned in the class
- **20**%: Program functionality and quality (e.g., modularization, clarity of variable names, program efficiency, documentation)
- **20**%: Quality of data manipulation and data insights, assessed both through the code and what is communicated in plain English by the group (orally and by text using slides/notebook/tables/figures/etc)
- **20**%: Effectiveness of the presentation: Clarity of the exposition, quality of slides (or Jupyter notebook markdown cells), respect of time allocation, validity of answers given to questions from the panel
- **10**%: Quality of graphical visualizations
- **10**%: Innovation and originality, concerning the project proposal, the final presentation, and the overall approach to data manipulation in Python

Forming Groups

- You are requested to work in a group of 2 (minimum) to 4 (maximum), and responsible for forming/joining a group
- The instructional team will create a Slack channel mid-September with instructions to form groups (in short: If you don't have a group yet, just post a brief introduction of yourself or your project interests; then reply in thread to other people's posts or DM them)
- Groups need be formed and final by October 18th: the instructional team will release an online form early October where each group will indicate a name for their group and the names of its group members

Project Proposal

- **Due November 2th by 8pmET**
- **Build your own proposal**: See list of recommended topics in section below. You can propose any original idea that involves data manipulation and/or analysis. Examples of proposals from last year will be presented on October 14th in a lecture dedicated to show you some project proposals. Note we are not authorized to distribute the PDF, so just please attend the lecture
- **Not graded**: The proposal itself is not graded directly, but can be taken into account during the final presentation when grading *innovation and originality*
- **Proposal format**: 2 to 4 slides. There is no requirement on who in the group present(s) the proposal. Presentation of the proposal itself is optional
- **Optional in-person presentation**: Each group is invited to present 5 min in front of a panel (DS-GA 1007 instructional team) on November 4th (location: 12 Waverly Place, Room G08). This presentation is informal, not graded, and optional. The goal is to benefit you = to receive feedback and advices from us, in particular on whether your proposal appears overly or insufficiently complex. You can also ask us questions about your project (total 10 min including Q&A)

Final Project

- **Due December 4th by 7pmET**
- **Code**: Complete (final) code must be submitted, together with a presentation (draft), by December 4th by 7pmET
- **In-person presentation**: December 6th (location: CDS, 7th floor Open Space) and December 9th (location: CDS, 7th floor Open Space), betwen 4pm and 9pmET, in addition to your group's assigned time slot, you can also attend presentations from other groups
- **In-person presentation**: Every group will present in front of a panel (DS-GA 1007 instructional team). This presentation is formal, graded, and mandatory
- **Presentation format**: Slides, or Jupyter notebook markdown cells instead of slides. No limit on number of slides (annexes welcomed too). Length of the presentation is limited based on group size: up to 6 min for groups of two, up to 8 min for groups of three, up to 10 min for groups of four. Everyone in the group must present (at least 2 min per person). The panel may ask questions to the group for up to 10 min after the presentation

Recommended Topics

- Historical descriptive analysis, or time series trends analysis (e.g., spread of epidemics, stock prices, taxi traffic patterns, product adoption, etc)
- Processing and/or statistical analysis of large data sets
- Mathematical models, scientific computing, machine learning
- Original visualization of data
- Social Network analysis, Natural Language Processing, etc
- Any data science project that lets you demonstrate your understanding of key Python concepts learned in the class

Clarification about Machine learning in this course: ML is not required in any form and is not the focus of the project assessment. The focus is on *programming* for data science. You are of course welcome to use ML models, but the panel will not grade the modeling part of your project