COURSE NUMBER: APSTA-GE.2017

Course Title: Advanced Topics in Quantitative Methods: Educational Data Science Practicum

Number of Credits: 2

Meeting Pattern: 3 hours per week, 7 weeks; first class meets Week of March 23, 2015.

Course time: TBD

Instructor: Y. Bergner

Course Description (~250 words or less):

This intensive laboratory course will focus on doing data analysis projects with real data selected by the students. The core skills are oriented around first framing good research questions, then having these guide interacting with data of all types and varying quality (e.g., web-scraped, or clickstream-based rather than large national surveys) via visualization, principled modeling and evaluation of models using statistical learning techniques such as regression, classification and clustering, and presentation of results, using "reproducible research" tools (e.g., knitr, sweave) in the R programming language.

Course Notes:

- Class sessions will be hands-on, involving pair programming in R and short student presentations, leading up to a final project. It will be a flipped class, in the sense that most information download (readings, watching videos, and interactive tutorials) will take place outside of class, lecturing will be minimal, and meetings will be dedicated to discussion and hands-on work. Students are required to be fully prepared to engage interactively in lab every week. This is not a survey course on applied machine learning algorithms, but rather a deep dive into the holistic process of knowledge extraction.
- Students deciding between APSTA-GE 2110 (Applied Statistics: Using Large Databases in Education) and this course should note the following.
    - APSTA-GE 2110 uses large national surveys (e.g., ECLS-K, NELS), teaches STATA software scripting skills (e.g., merging databases), and emphasizes best practices in applied statistical analysis of survey data (e.g., weighting and adjusting standard errors in complex surveys). The course "pairs" nicely with RESCH-GE 2139 (Survey Research) PADM.GP.2902 (Multiple Regression and Introduction to Econometrics).
    - This course, APSTA-GE 20xx, may use "messy" (non-tabular) data that is more common to web-derived data, such as clickstreams and web-scraping (processing the text in HTML files), but the type of data may be tabular as well. The tools used in this course are different and the skills needed to use these tools emphasize computer science skills in addition to statistical modeling skills, which also differ (APSTA-GE 2011 (Classification and Clustering) is a prerequisite). Thus, students are expected to have some prior programming experience and will quickly adapt these to learn the R programming

language as well as tools that integrate computer code, statistical analysis, and technical writing (e.g., the knitr library).

- We recommend this course to advanced PhD, MS-A3SR and MS Data Science students, provided they have prior programming experience and the appropriate prerequisite knowledge.

Course Prerequisites/Expectations:

- Some prior experience with statistics (e.g., APSTA-GE 2003) and some exposure to programming (experience with loops, functions, and data structures, e.g., at the level of NYU's undergraduate course "Introduction to Computer Science (CSCI-UA-0101)) in any language, e.g., C, Java, Python, R);
- Students are strongly encouraged to take APSTA-GE 2011, Classification & Clustering –offered Winter 2015—before taking this course (the classification and clustering topics are central to this course).
- Students are also encouraged to use the first half of the Spring 2015 term leading up to this course gathering data sources and considering research questions which the data are intended to address.

Learning Objectives:

By the end of the course, students will be able to:

1. Formulate, contextualize, and defend research questions.
2. Explore and transform educational data sets using R.
3. Understand how regression, classification and clustering methods are applied to educational data and how to evaluate these methods.
4. Generate high quality visualizations.
5. Present methods and results according to professional standards of reproducible research.
6. Learn from and collaborate with others on educational data science projects.

Course Format: (Lecture, lab, seminar, recitation or combination)

One combined lecture and lab session each week; Spring offering

Course Outline (list of lectures/topics each session)

Week	Topics	Lab Activity	Advance Preparation
1	Overview; R, knitR, and swirl; finding and reading data files	Data manipulation visualization	Bring a laptop to class!
2	Descriptive statistics; cleaning data; transforming variables; missing data	First 5 minute presentation; descriptive statistics lab	Swirl "R Prog Alt", parts 1-5. External data mini-project.
3	Regression-based methods	Regression trees; Second chance	Strobl et al.

		presentations (SCPs)	
4	Classification-based methods; association measures for categorical variables	Midterm project presentation; classification lab	Read vcd-Tutorial (on CRAN)
5	Clustering methods; internal and external quality indices	Clustering lab; SCPs; project time	Steinley
6	Visualization	Viz lab; SCPs; project time	Tufte; Gelman
7	Presentation and peer review	Final presentations (Peer reviews due within 1 week)	Peer review protocol

Course Requirements

There will be 3 projects, presented and written up individually. Students are not only encouraged to work together, but each time you will be assigned a partner as an "internal" reviewer. Furthermore, students will write "external" peer reviews for one midterm and one final project. Projects will be scored on a four-point scale: {strong accept, weak accept, weak reject, strong reject}, and you will not be considered eligible to present the next project until you *and your assigned partner* have both received strong accepts on your prior presentations (see grading notes below). Second chances are of course permitted.

Evaluation for this course will be weighted as follows:

- Projects
    - Mini-project 10%
    - Midterm project 20%
    - Final project 40%
- Peer reviews 10%
- Class participation 20%

ASSIGNMENT AND GRADING DETAILS

Projects:

The three projects will be assessed for excellence in: quality of the code (well-commented; functional); organization of the code/writing (is the research question clearly stated? Are the data well described? Are the analytic techniques clear and well executed? Is the flow of the presentation well-organized); and reproducibility/flexibility/extendibility of the code (how modular is the design? Could the structure be reused for a slightly different problem?). To receive maximum credit for each project, satisfaction of all three requirements is required.

Reviews:

Final and midterm projects will be both peer reviewed (rubric will be provided in class) and rated by the instructor on the same four-point scale of {strong accept, weak accept, weak reject, strong reject}. To emphasize the importance of writing detailed peer reviews, student written reviews will be scored on the scale: 1=inadequate, 2=useful, 3=exemplary.

Class Participation:

This course is highly interactive, both in terms of working and learning in teams and as a classroom. However, interaction takes a variety of forms, ranging from one-on-one discussions to group presentations, so that different skills are emphasized at different times. The evaluation of class participation uses a flexible scale so that everyone can achieve the highest measure. For each class meeting, 1=present, 2=responsive, 3=active, and the overall participation grade is obtained by summing over the class sessions.

The following system can be used to convert evaluation scales used in this course to letter grades:

Letter grade	**Projects**	**Peer review**	**Class participation**
A	Strong accept	Exemplary	Active
B	Weak accept	Useful	Responsive
C	Weak reject	Inadequate	Present
D	Strong reject		

Required Readings and/or Text (a partial reading list is acceptable)

Gelman A. (2013). Choices in statistical graphics: My stories. Presentation to New York Data Visualization Meetup. URL: http://www.stat.columbia.edu/~gelman/presentations/vistalk_meetup_new_handout.pdf

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59,* 1-34.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Tufte, E. R. (2001). Graphical Excellence. In *The Visual Display of Quantitative Information* (pp. 13–51).

There will be a number of readings, videos, and tutorials available from the web.

Academic Integrity:

All students are responsible for understanding and complying with the NYU Steinhardt Statement on Academic Integrity. A copy is available at: http://steinhardt.nyu.edu/policies/academic_integrity.

Students with Disabilities:

Students with physical or learning disabilities are required to register with the Moses Center for Students with Disabilities, 726 Broadway, 2nd Floor, (212-998-4980 and online at http://www.nyu.edu/csd) and are required to present a letter from the Center to the instructor at the start of the semester in order to be considered for appropriate accommodation.