DEPARTMENT OF STATISTICS AND DATA SCIENCE

Communicating Quantitative Analyses

STAT 4020 Spring 2023

Professor Richard Waterman, 315 WARB, waterman@wharton.upenn.edu

Office hours: M/W 1:30pm – 2:30pm.

Class meets: MW 12:00pm – 1:30pm. Room 302 ARB (conference room).

Overview

This seminar-based capstone course provides an opportunity for students to hone their data science and statistical modeling skills, together with an emphasis on communicating quantitative results. This is not a "theoretical class", but rather, experiential. It allows students to bring their existing knowledge from different disciplines to bear on new problems. Four real-life datasets will be analyzed during the quarter, and students will be expected to create and deliver in-class presentations for each analysis. The goal is to create and share compelling insights, which is not always the same as using the most advanced techniques possible. What makes this class different from other analytics classes is the emphasis on the public presentation of results. The more you stand up in front of people to explain your ideas, reasoning, and conclusions, the better you get at it. There is no substitute for experience. This class will give that experience.

It is expected that students will proactively investigate new methodologies and graphical techniques as the homeworks require. Identifying and using new styles of graphs (not just histograms and boxplots) is strongly encouraged.

Audience

The course will be suitable for anyone who wants more opportunities to analyze data, continue developing their programming skills and those who want to gain experience and confidence in presenting results and conclusions to an audience.

Pre-requisites

The course presumes that students have taken a sequence of stat courses such as STAT 1010/1020, or 4300/4310 and so are familiar with multiple regression analysis. In addition, they should have been exposed to more advanced techniques such as logistic regression and tree-based methods as taught in classes like STAT 4220/4230/4710. Finally, it will be assumed that students have knowledge of a programming language such as R or Python and an IDE such as R-Studio or Jupyter notebooks. Classes such as STAT 4050/4700 would meet this requirement.

If you have taken similar courses outside of the Statistics Department, please contact the instructor to discuss their suitability as replacements.

Format

The class will meet twice a week each time for 1.5 hours in a seminar format. During class time we will discuss the datasets, review potential graphical techniques and methodological approaches, listen to group presentations, and provide feedback.

Homeworks will be done in randomly created groups of four, and there will be a structure to provide one another with feedback.

Because of the collaborative nature of the class, **attendance is mandatory**.

Course materials

Lecture notes and topical papers available via Canvas.

Software

The software used in the course can be R, Python or JMP as appropriate.

Books

There is no required book.

Homeworks

There will be four homeworks. Each homework has three components.

1. A code file(s) used to create the content of the presentation.

2. Presentation slides saved as a PDF or reveal.js file.
3. The in-class delivery of the presentation including a Q&A component.

HW 1: EDA and descriptive statistics. Mining injuries.

The first homework involves analyzing a dataset using only descriptive and graphical tools. Not everything needs to be modeled and often the most compelling insights can be gained through a careful graphical analysis. The dataset contains the number of injuries at mines within the US. Features of each mine are included such as the number of employees, the commodity being produced etc.

HW 2: Continuous Y-variable. Bike rentals.

The second dataset contains data on the number of bike rentals for each hour of the day in a city. Weather conditions, time of year, weekday, holiday status are provided, and the aim is to create a predictive model of the number of rentals expected in any hour.

HW 3: Categorical Y variable. Predicting high/low value customers.

This dataset has a categorical outcome that buckets insurance policy holder into two groups; high value and low value based on profitability. Explanatory variables such as age, education, marital status etc. are provided as potential predictors.

HW 4: Student chosen group project.

For the final homework students will identify a publicly available dataset and frame a related business problem. They will then perform a self-directed analysis which each group will present to the class. We will spend class reviewing the selected datasets and discussing/brainstorming each group's plans.

Late homework policy: moot. Because the homework grade is based on the project presentations, not presenting would result in a 0 for that project.

Grading

The final course grade will be made up of:

Attendance: Mandatory. (10%).

Class participation (10%).

HW 1 (10%: 3% for the group presentation, 2% individual presentation, 5% intra-group feedback).

HW 2 (20%: 10% for the group presentation, 5% individual presentation, 5% intra-group feedback).

HW 3 (20%: 10% for the group presentation, 5% individual presentation, 5% intra-group feedback).

HW 4 (30%: 15% for the group presentation, 5% individual presentation, 10% intra-group feedback).

Course timetable

Class	Activity
1. 1/11/2023	Intro to the course and HW 1 discussion
2. 1/18/2023	HW 1 discussion
3. 1/23/2023	HW 1 presentations
4. 1/25/2023	HW 2 discussion
5. 1/30/2023	HW 2 discussion and Poisson regression
6. 2/1/2023	HW 3 discussion and modeling ideas
7. 2/6/2023	HW 2 presentations
8. 2/8/2023	HW 3 discussion
9. 2/13/2023	HW 3 discussion and HW 4 intro
10. 2/15/2023	HW 3 presentations
11. 2/20/2023	HW 4 discussion
12. 2/22/2023	HW 4 discussion
13. 2/27/2023	HW 4 presentations
14. 3/1/2023	Reflection and wrap-up. Asynchronous.