ANLY 506: Exploratory Data Analysis

Olga Scrivner, PhD

"Three of the main strategies of data analysis are: 1. graphical presentation. 2. provision of flexibility in viewpoint and in facilities, 3. intensive search for parsimony and simplicity."
(Jones, 1986, Vol. IV, p. 558)

Course Description

Exploratory data analysis plays a crucial role in the first stages of analytics. It comprises the pre-processing, cleaning, and preliminary examination of data. This course provides instruction in all aspects of exploratory data analysis. It looks at a wide variety of tools and techniques for pre-processing and cleaning data, including big data. It provides students with practice in evaluating and plotting/graphing data to evaluate the content and integrity of a data set.

Course Objectives

At the end of this course students will:

- Understand the intellectual foundations for EDA and information organization.
- Understand and be able to implement tools and techniques related to EDA and information retrieval from a wide variety of sources.
- Be able to pre-process, clean and evaluate data.
- Be able to perform basic data visualization tasks.
- be able to construct and interpret a variety of data analyses.

Required Textbooks

Phillips, N. D. (2016). Yarrr! The pirateâĂŹs guide to R. Available at https://bookdown.org/ndphillips/YaRrr/

Peng, R. D. (2015). The Art of Data Science: A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC. Available at http://leanpub.com/artofdatascience

Grolemund, G., Wichham, H. (2018). R for Data Science. Available at https://r4ds.had.co.nz/index.html

Prerequisites

Prerequisites: R Basics

Course Structure

Weekly lectures, quizzes, reading, hands-on R practice, code portfolio and 2 exams.

Quizzes

Each week you will have a quiz based on readings and lectures. The hard deadline for submission is on Sundays 11:55pm EST.

Exams

There will be one midterm exam and one final exam based on the theory learned from lectures and applied knowledge learned from practice.

Grading Policy

No late assignments are accepted unless an official documentation is provided (e.g., medical documentation). Business trips, job interview, vacations are not considered as valid excuses. Students who participate in University-sanctioned events (such as athletics) must make prior arrangements and give the instructor at least one week notice.

- **40%** 2 Exams
- **10%** Assignment
- **5 %** Attendance
- **15 %** Code Portfolio
- **30%** Quizzes

Per HU policy: The Grad School valid grades are A, A-, B+, B, B-, C+, C, and F.
Grading scale: A=90-100; A-=88-89; B+=85-87; B=82-84; B-=79-81; C+=75-78; C=73-74; F<73

Statement on Academic Integrity

According to the University's Student Handbook: Academic integrity is the pursuit of scholarly activity free from fraud and deception, and is the educational objective of this institution. Academic dishonesty includes, but is not limited to cheating, plagiarism, fabrication of information or citations, facilitating acts of academic dishonesty by others, unauthorized possession of examinations, submitting work of another person, or work previously used without informing the instructor, or tampering with the academic work of other students. Any violation of academic integrity will be thoroughly investigated, and where warranted, punitive action will be taken. Students should be aware that standards for documentation and intellectual contribution may

depend on the course content and method of teaching, and should consult the instructor for guidance in this area.

Honor Code - We as members of Harrisburg University community pledge not to cheat, plagiarize, steal, or lie in matters related to academic work. As a Community of Learners, we honor and uphold the HU Honor Code.

Schedule

The schedule is subject to change.

Week	Theory	Practice
Week 1	Introduction to Course Introduction to EDA Read: Page 1-9 Tukey, John W. (1977). Exploratory Data Analysis. Practice: Review R basics with DataCamp https://www.datacamp.com/courses/free-introduction-to-r	quiz 1
Week 2	EDA Data and Workflow Read 1: Chapter 2 from Peng and Matsui (2018) Workflow. Read 2: Data Taxonomy in R http://www.r-tutor.com/r-introduction/basic-data-types Read about Numeric, Integer, Logical, Character	quiz 2
Week 3	Data Import/Export Read: Chapter 3 Stating and Refining the question (Peng & Matsui, 2016) Practice: Chapter 9 (Phillips, 2016). Yarrr! The pirate's guide to R. https://bookdown.org/ndphillips/YaRrr/importingdata.html	quiz 3
week 4	Data Structures Read: Chapter 8. (Phillips, 2016). Yarrr! The pirate's guide to R. Practice: Chapter 8.7 (Phillips, 2016)	quiz 4
week 5	Data Transformation Read: Section 1 and 2 Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1-23. https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf	quiz 5
week 6	Practical Stats Read: Bruce PC, Bruce A (2017) Practical statistics for data scientists. Practice: Chapter 5 https://r4ds.had.co.nz/transform.html	quiz 6
week 7	RMarkdown and Workflow	Quiz 6
week 8	Exploratory Visualization I	Exam 1
week 9	Spring Break II	
week 10	Exploratory Visualization II	quiz 7
week 11	Cluster Analysis	quiz 8
week 12	Factor Analysis	quiz 9
week 13	Time Series Analysis	quiz 10
week 14	Dimension Reduction	quiz 11
week 14	Big Data	quiz 12
week 15	Advanced EDA	quiz 13