Data Science Principles and Practice - UC Berkeley Extension

Computer Science X415.1 (Classroom Course - 2 semester units in Computer Science)

Wednesdays, Feb. 8 to April 12, 6:30 PM - 9:30 PM

Course Canvas web site: https://onlinelearning.berkeley.edu/

(**Bring your laptop to every class meeting**)

Instructor:

Allan Miller (a.k.a. Allan, Professor Miller, Dr. Miller - your choice)

allan.m.miller@berkeley.edu

Mobile: (510) 326-1486

Course Description:

This course provides a hands-on, practical introduction to Data Science using tidy data principles and the R programming language. The course covers:

- The Data Science process
- The tidy approach for collecting, preparing, exploring, and managing data for analysis
- Exploratory Data Analysis (EDA)
- Data Visualization: advanced graphics and interactive, web-based analytics applications
- Introduction to building and evaluating models, presenting results

Course Objectives:

When you have completed this course, you will know how to:

- Understand the roles, stages and importance of deliniating the goals of a Data Science project
- Utilize advanced, tidy data manipulation methods to access, transform, and load data into R for analysis
- Conduct rigorous data exploration including the use of advanced data visualization techniques to understand individual variable distributions and explore relationships between variables, identify issues with data outliers, units, data types, missing, and invalid values
- Prepare data for analysis by fixing data quality problems, tranforming data for the modeling process.
- Develop interactive data visualizations
- Understand the basic principles for choosing appropriate modeling techniques
- Understand the basic principles for evaluating the effectiveness of models, identifying and remediating model problems
- Effectiveley present model results, including effective visualizations
- Understand the importance of, and know how to use methods for doing reproducable research.

Prerequisites:

- Programming With R EL ENG X480.1 or, equivalent knowledge of R:
    - Familiar with and experience programming base R data structures, creating, manipulating, and accessing: vector, matrix, factor, data frame, and list
    - Familiar with, and experience using, R for vectorized programming and common fundamental R data types (numeric, character, logical)
    - Familiar with commonly used base R functions such as mean, mode, rnorm, runif, log, is.na, etc..
    - Know how to install and load R packages, write and call programmer-defined R functions, use R to solve basic data manipulation and computation problems
    - Familiar with the RStudio development environment including knitting basic RMarkdown documents to HTML
    - Proficient in base R and ggplot2 graphics
- Basic knowledge of statistics as covered in a first-semester undergraduate statistics course.

Intended Audience:

- Data Analysts who wish to learn and use advanced Data Science techniques in their work
- Students who wish to prepare for a career in Data Science
- Managers and Engineers who wish to pursue a career in Data Science
- Scientists, engineers, business analysts, and social science researchers who want to apply Data Science principles and practics in their work

Course Topics, Readings, and Schedule (subject to change):

Topics Summary:

Part 1 (Weeks 1 - 4): tidy data principles for loading, transforming, and exploring/visualizing data

Part 2 (Weeks 4 - 6): applying tidy data principles to accessing, loading, tranforming, exploring and visualizing categorical, time series, and geospatial data sets

Part 3 (Weeks 7 - 10): Modeling basics

Readings:

Note:
R4DS = R for Data Science
PDS = Practical Data Science

Weeks 1-4

- Course introduction
    - R4DS Preface

    - R4DS Chapter 6 *Workflow: projects*
    - PDS Chapter 1 *Introduction to Data Science*
- The tidy approach to data manipulation (skim):
    - *Tidy Data*, Wickham, **Journal of Statistical Software**, Vol 59 (4014), Issue 10
    - R4DS Chapter 3 *Data transformation with dplyr* (advanced coverage)
    - R4DS Chapter 14 *pipes with magritr*
    - NSE Non-standard evaluation
- Exploratory Data Analysis (EDA):
    - R4DS Chapter 5 *Exploratory Data Analysis*
    - PDS Chapter 3 *Exploring Data* (up to but not including Bar Charts)
    - PDS Chapter 4 *Managing Data* (through 4.1, incluing missing data, omit sampling for modeling and validation)
- Wrangle data:
    - R4DS Chapter 7 *Tibbles with tibble*
    - R4DS Chapter 9 *Tidy data with tidyr*
    - R4DS Chapter 12 *Factors with forcats*
    - vignette: *Working with categorical data with R and the vcd and vcdExtra packages*, Friendly

Weeks 4, 5, 6: Working with Categorical, Time Series and Geospatial Data

- Time series data
    - R4DS Chapter 13 *Dates and Times with lubridate*
    - TBA
- Geospatial data
    - *Introduction to visualising spatial data in R*, Lovelace, Cheshire
    - ftp://ftp.bgc-jena.mpg.de/pub/outgoing/mforkel/Rcourse/spatialR_2015.pdf, Forkel
    - Working with leaflet: interactive maps with R

Weeks 6, 7: Interactive data visualizations using Shiny

- RStudio: *Introduction to Shiny Tutorial*
- RStudio: *Shiny Articles*

Weeks 7, 8, 9 10: Model basics

- R4DS Chapter 18 *Model Basics with modelr*
- R4DS Chapter 19 *Model Building*
- R4DS Chapter 20 *Many Models with purr and broom*

- PDS Chapter 5 *Mapping problems to machine learning tasks*
- PDS Chapter 6 *Memorization Methods*
- PDS Chapter 7 *Linear and logistic regression*

Assignments (to be posted on Canvas):

- Assignment 1: preliminaries, using R
- Assignment 2: load, tidy, wrangle and transform data; visualize results
- Assignment 3: load, tidy, wrangle, and transform time series, geospatial data; visualize results using a shiny-based web application
- Final Project: application of tidy data science principles

Note: **no late programming assignments accepted**. Turn in assignments as specified, on time, to receive credit.

Instructional Methodology

- In-class presentation of main topics
- Work through examples to illustrate concepts
- In-class, hands-on exercises for students to work on and discuss (bring a laptop with R installed and running to every class meeting)
- Online message and discussion forum (Canvas: onlinelearning.berkeley.edu)

Credit Requirements (percentages of final course grade):

- Programming projects (20%)
- Final project due April 22 (30%)
- Final exam April 12 (35%)

  Short answer problems, writing some R code (similar to class exercises). **Final exam is closed book, open notes. You may use your class notes, exercises, and projects during the final exam** (printed or laptop copy). No use of web sites, online books, or R during the final exam.
- Class participation, including participation in class discussions, completion class exercises, and online forum participation (15%)

Grade Options:

- Letter Grade
- Credit/No Credit (earned grade C or above)
- Not For Credit (audit, assignments not required, submitted)

Note: letter grade is the default grade option. Notify me by email no later than March 29 (only) if you wish to change your grade option to Credit/No Credit, or Not For Credit.

Texts (required):

R for Data Science: Visualize, Model, Transform, Tidy and Import Data by Hadley Wickham, Garrett Grolemund, 2016 O'Reilly Media 1st edition
ISBN 9781491910399

Practical Data Science with R by Nina Zumel, John Mount, 2014 Manning 1st edition
ISBN 9781617291562

Computing Resources

- Access to a R and RStudio to complete programming assignments: http://www.rstudio.com/
- **Bring your laptop to every class session** with RStudio installed to work on classroom hands-on exercises.

Notes:

- In-class examples and exercises discuss model solutions to problems.
- There are no official class notes for the course. Code examples will be posted to Canvas or a public repository on GitHub:
- If you need to miss a class session, please let me know by email beforehand. Assignments and handouts will be posted on the course web page.
- A "Not for Credit" grade option is available if you wish to participate in the class, but are unable to attend class meetings, complete assignments, and take the final exam. Strongly recommended for students who have concerns about their academic preparedness for taking the class or outside of class work or personal commitments.
- All students must be officially enrolled through U.C. Berkeley Extension.