UNIVERSITY OF MICHIGAN
MEDICAL SCHOOL
DEPARTMENT OF LEARNING HEALTH SCIENCES

LHS 610
Exploratory Data Analysis for Health

CREDIT HOURS	3
PRE-REQUISITES	Graduate standing or permission of the instructor. Advisory pre-requisites: Students should have taken one course in statistics and have experience with either a statistical or general-purpose programming language.
CLASS SCHEDULE	Fridays 1:10-4 PM
LOCATION	Med Sci 2 – Room 2813/17 (except September 22, see below)
FACULTY	Karandeep Singh, MD, MMSc; V.G. Vinod Vydiswaran, PhD

COURSE DESCRIPTION

Real health data is complex, often unstructured, at times inaccurate, inconsistent, contains missing values, and is organized for clinical care rather than to meet analytic needs. Learning from health data requires a solid grasp of data operations, data visualization, statistics, and machine learning, as well as an understanding of ethical and legal frameworks guiding health data privacy and security. Students in this course will learn foundational topics in data science focused on health data and will apply this knowledge on real health datasets through hands-on labs integrated into the lectures. The course is based on two large themes: (a) understanding health data, and (b) making inferences based on data. Students will develop a systematic working understanding of R, one of the most widely used languages for data science, and an introductory understanding of several packages useful in analyzing health data. They will participate in a group project focused on answering a health-related question. After completing this course, students should be able to securely store a health data set, summarize its structure, merge tables, visualize relationships, reshape and subset it to meet analytic needs, deal with missing values, apply statistical and machine learning methods to build prediction models, and evaluate the performance of these models.

COURSE OBJECTIVES

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

1. Define the challenges of working with health data, including legal and methodological issues pertaining to the secondary use of clinical data.
2. Describe the features of a health-related question that make it important and answerable, including data sources that can be leveraged to address it.
3. Apply “data verbs” and visualization to identify data quality issues and describe important relationships between clinical features and health outcomes.
4. Apply machine learning methods to develop and validate predictive models to answer a health-related question.
5. Discuss strategies for working with big data and speeding up data analysis.

TEACHING METHODS

This course will be taught using multiple methods, including lecture, in-class laboratory exercises, assignments, a mid-semester group project, and a final group presentation.

GRADING AND EVALUATION

Instructors will facilitate in-class laboratory exercises and provide mentorship for the course projects.

The course evaluation will be based on:

- Homework assignments (75%)
- A mid-semester group project on summarizing and visualizing data to answer a health-related question (10%)
- A final group project and presentation on building a machine learning model to answer a health-related question (15%)

WEEKLY SCHEDULE

Prior to week 1: Install R and RStudio

Week 1: Introduction and R basics (September 8, 2017)
- Packages: readr (part of tidyverse)
- Lecture:
    - Introduction, motivation for course
    - Why R?
    - Challenges in working with health data
        - Primary versus secondary use of data
        - Identified data, limited use data, de-identified data
        - HIPAA privacy rule, HIPAA security rule
    - Basics of R and RStudio
        - Introduction to tidyverse
        - Reading in data from CSV
        - Univariate statistics – median, mean, minimum, maximum, standard deviation, interquartile range
        - Installing packages
        - Accessing help
    - Course administration issues (office hours, homework, projects, grades)
- Exercise:
    - Working with R notebooks for homework assignments
- Homework: Assignment 1 available

Week 2: Data structures, data frames, and more descriptive statistics (September 15)
- Packages: dplyr, readxl, haven, desctable
- Lecture:
    - Reading in data from other sources in R
        - Excel, SAS, SAS export, and STATA files
    - Data structures
        - Vectors, matrices, lists, data frames
    - Data frame column types
        - Logical, integer, numeric, characters, factors, dates
    - The verbs of data analysis
        - Selecting, subsetting (filtering/slicing), mutating, summarizing, grouping
    - Grouped univariate statistics and generating a Table 1
- Exercise:
    - Read in NHANES blood pressure data set
    - Describe the blood pressure of the U.S. population
- Homework: Assignment 1 due, Assignment 2 available

Week 3: What makes a health-related question important and answerable? (September 22)
- **NOTE: CLASS IS IN THSL 2903**
- Packages: none.
- Lecture:
    - What makes a health-related question important?
        - The “triple aims” of health care

        - Has the question already been answered?
    - What makes a health-related question answerable?
        - Hill's criteria for causation
        - Having the (right) data available
        - Sources of bias in study design
        - Sample size and power
- Exercise
    - Form groups and brainstorm ideas for midterm projects
        - Use PubMed and other research and news sites to determine importance of questions
        - Search the web for datasets that may be able to answer that question

Week 4: Joining and reshaping health data (September 29)

- Packages: dplyr, tidyr
- Lecture:
    - Imperfections of real-world data – what makes a dataset "messy" or "tidy"?
    - Recoding variables with if_else, case_when
    - Joining data frames
        - Left join, right join, inner join, outer join
    - Reshaping data
        - Long to wide
        - Wide to long
- Exercise:
    - Use NHANES to answer these questions: How many medications does the average person take? What proportion of the population takes both aspirin and acetaminophen?
- Homework: Assignment 2 due, Assignment 3 available

Week 5: Visualizing health data (October 6)

- Packages: ggplot2
- Lecture:
    - Principles of visualization
    - Grammar of graphics
    - How to use ggplot2
    - Visualizing 2-dimensional relationships
        - Continuous vs. continuous variables
        - Categorical vs. continuous variables
        - Categorical vs. categorical variables
    - Adding more dimensions to the plot
    - Mid-semester course feedback
- Exercise:
    - What is the relationship between blood pressure and age?
    - What is the relationship between blood pressure and kidney disease?
- Homework: Assignment 3 due, Assignment 4 available

Week 6: Hypothesis testing (October 13)

- Packages: none.
- Lecture:
    - The null hypothesis and the alternative hypothesis
    - What are p-values and how should we interpret them?
    - Rethinking visualizations as statistical tests
        - Continuous vs. continuous variables (correlation)
        - Categorical vs. continuous variables (difference in means/rank)
        - Categorical vs. categorical variables (difference in proportions)
    - Testing multiple hypotheses without engaging in "p-hacking"

- Exercise:
    - What is the relationship between blood pressure and age?
    - What is the relationship between blood pressure and kidney disease?
- Homework: Mid-term project available

Week 7: Supervised machine learning 1 (October 20)

- Packages: mlr
- Lecture:
    - Concepts in supervised machine learning
    - Rigid vs. flexible models
    - Decision trees
    - K-nearest neighbors
    - Perceptron
    - Support vector machines
- Homework: Assignment 4 due

Week 8: Supervised machine learning 2 (October 27)

- Packages: mlr
- Lecture:
    - Introduction to mlr
    - Evaluating out of sample performance
        - Training set and validation set
        - Cross-validation
            - K-fold and leave-one-out
        - Bootstrapping
    - Performance metrics
        - No information rate
        - Sensitivity, specificity, positive predictive value, negative predictive value
        - Precision, recall, and F-measure
        - Kappa statistic
        - The receiver-operator curve (and area under the curve)
    - Introducing the final project and expectations for the *proposal*
- Exercise:
    - Train a machine learning model to predict whether a patient has breast cancer based on pathology findings
    - How well would this model be expected to perform on a new set of data?
- Homework: Mid-term project due, Final project *proposal* available

Week 9: Handling missing data (November 3)

- Packages: mlr
- Lecture:
    - How does R treat missing data?
    - Analysis of missing data
        - Complete case analysis
        - Imputation
            - Mean, median, mode, histogram
            - Model-based imputation
            - Multiple imputation
- Exercise:
    - Impute missing values of blood pressure in NHANES. How does the distribution of blood pressure change based on if you use mean, median, and model-based imputation?
- Homework: Assignment 5 available

Week 10: Supervised machine learning 3 (November 10)

- Lecture:
  - Challenges with flexible models
    - Overfitting
    - Non-informative features
  - Potential solutions
    - Feature selection
    - Ensemble methods
      - Bagging
      - Boosting
- Exercise:
  - Group machine learning activity: “To understand ensemble methods, you must become the ensemble.”
- Homework:
  - Assignment 5 due, Assignment 6 available
  - Final project *proposal* due, Final project available

Week 11: Strategies for dealing with big data (November 17)

- Packages: dbplyr, mlr
- Lecture:
  - Strategies for storing big data
    - Sparse representation
  - Strategies for reading in big data
    - Reading from databases
  - Strategies for speeding up data analysis
    - Understanding bottlenecks in code performance
    - Parallel processing
- Exercise:
  - Share your 1-minute pitch for final project proposal
  - Run multiple mlr models in serial and in parallel and compare times – which is faster?
- Homework: Assignment 6 due

Week 12: Analyzing text data (December 1)

- Packages: stringr, tidytext
- Lecture:
  - Reading text data into R
  - Word and document frequency
  - N-grams
  - A primer on regular expressions
- Exercise:
  - What is the most frequent word in a set of health notes?
  - What is the most common bigram in a set of health notes? How about trigram?

Week 13: Final group project presentations (December 8)

- **NOTE: CLASS WILL BE 12-4 PM**
- Final group project presentations
- Homework: Final project *slides* due, Final project *follow-up questions a*ssigned

Week 14: No class (December 15)

- Homework: Final project *follow-up question* answers due

REQUIRED TEXTS

R for Data Science by Garrett Grolemund and Hadley Wickham

- Available online: http://r4ds.had.co.nz

- Also available in print

Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

- Available online: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf
- Also available as an e-book through Mirlyn (http://mirlyn.lib.umich.edu)
- Also available in print

Additional readings will be on the course Canvas site.

COURSE POLICIES

1. Late submission policy

Students have 72 hours of buffer grace period for the entire semester. If necessary, students may use it to submit any of the assignments, homework, or the course project reports late without any effect on the overall grade. The grace period, however, cannot be used on the student presentations and the final project. A student may use it all on one assignment or use a bit of it for any number of assignments. Once the buffer grace period is used up, late submissions will not be graded.

2. Academic Conduct

2.1. Collaboration

The instructors strongly encourage collaboration while working on homework assignments. Collaboration with other students in the course will be especially valuable in learning programming skills. To further support collaboration, the course Canvas page will host a forum where students may answer one another's questions. You must write your own code and narrative answers, and homework assignments must represent your own work. If you worked with someone on the homework before writing it, you must list any and all collaborators on your written submission. You may not share blocks of code or narrative answers to homework questions with other students. For group projects, sharing of code and answers between members of the same group is expected. However, students in different groups may not share blocks of code or narrative answers. Read the instructions carefully and request clarification about collaboration when in doubt.

2.2. Plagiarism

All written submissions must be your own, original work. Original work means you may not simply modify someone else's completed code or narrative answer; these must not be shared between students, except between members of the same group for group projects. You may incorporate code from class exercises into your assignments; this does not require attribution. You may obtain copy editing assistance, and you may discuss your ideas with others, but all substantive writing and ideas must be your own, or be explicitly attributed to another. See the Rackham Academic and Professional Integrity Policy (http://www.rackham.umich.edu/current-students/policies/academic-policies/section11) for the definition of plagiarism, resources to help you avoid it, and the consequences for intentional or unintentional plagiarism.

3. Reasonable accommodations

The university will provide reasonable accommodations to qualified individuals with disabilities upon request. If you think you need an accommodation for a disability, please let the instructors know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. The instructors will treat any information that you provide in as confidential a manner as possible. For more information, see https://ssd.umich.edu/article/americans-disabilities-act-ada

It is also the University's policy that every reasonable effort be made to help students avoid negative academic consequences when their religious obligations conflict with academic requirements. Students who expect to miss classes, examinations, or other assignments as a consequence of their religious observance are requested to contact the instructors by the drop/add deadline. For more information, see https://www.provost.umich.edu/calendar/religious_holidays.html