Introduction to Data Science

Introduction

We have entered a time in which vast amounts of data are more widely available than ever before. At the same time, a new set of tools has been developed to analyze this data and provide decision makers with information to help them accomplish their goals. Those who engage with data and interpret it for organizational leaders have taken to calling themselves data scientists, and their craft data science. Other terms that have come into vogue are *big data*, *predictive analytics*, and *data mining*. These can seem to be mysterious domains. The point of this class is to demystify much of this endeavor for individuals who will be organizational leaders.

The class is structured around developing students' skills in three areas: getting data, analyzing data to make predictions, and presenting the results of analysis. For each area, the subtopics are as follows:

Getting Data Topics

- Tools of the trade: R and RStudio
- Working with pre-processed data and flat files
- Getting data from the web: webscraping, using forms, using application programming interfaces
- Using databases

Analyzing Data Topics

- Descriptives and conditional means
- Regression
- Supervised learning: classification
- Unsupervised learning: *K*-means and nearest neighbors clustering
- Cross validation

Presenting Data Analysis Topics

- Descriptives: histograms, density plots, bar plots, dot plots
- Scatterplots
- Lattice graphics and small multiples
- Interactive graphics
- Communicating results effectively

Evaluation

Students will be evaluated based on two areas: weekly problem sets and the final project.

- Problem sets 65%: Each week students will be assigned a problem set to complete. The problem sets will be due 24 hours prior to the following week's live session. For example, the Week 1 problem set will be due 24 hours prior to the Week 2 live session. No late problem sets will be accepted. Each problem set will be graded on a 100-point scale. Your lowest grade will be dropped.

  There will be 13 assigned problem sets, with each problem set worth 100 points. The lowest grade will be dropped, meaning that you will be graded on 12 of these problem sets. The grading standards will be as follows:

  50 = turned in problem set, did not attempt most of the problems

  75 = turned in problem set, attempted most of the problems

  100 = turned in problem set, attempted all of the problems

  Note: All HW Problem Sets Submissions must be in "knitted" format: html, doc, or pdf. There may be a penalty for submissions not adhering.

  Note that your grade on problem sets does not depend on your being correct on all problems but simply making a serious attempt to answer all of the problems.

- Final Project 35%: During the course of the semester you will work on a final assignment utilizing your skills as a data analyst.
  - Progress reports 17.5%: 100 points each
  - Final Product 17.5%: 100 points

Texts

Required Texts

We will have two texts for the course. The first is Hadley Wickham's book, R for Data Science:

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. San Francisco, CA: O'Reilly Media, Inc.

The other text is Nate Silver's *Signal and the Noise*:

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—but some don't*. New York, NY: Penguin.

Software

We will use only free, open-source software in this course.

We will use R, an open-source data analytic platform for all analysis. R appears to be the most widely used data analysis software in data science. We will utilize RStudio as our integrated development environment (IDE) for R.

Honor Code Statement

All assignments for this class, including weekly problem sets and the final project, are to be conducted under the obligations set out in Vanderbilt's Honor Code. Please click here to review the Honor Code.

There will be two quite different standards for completing the problem sets and the final project.

Problem sets. You may collaborate with anyone, and you may utilize any resource you wish to complete these problem sets.

Final Project. All of the work on the final assignment must be your own. Anyone's work that you reference should be cited as usual. All data that you do not personally collect must be cited, as with any other resource.

If you have any questions at all about the Honor Code or how it will be applied, ask me right away.

Schedule

Weekly Schedule:

We will be deviating from the LMS Module Sequence of topics for the course. The goal is to provide a more contiguous and cohesive flow to the course and list of topics.

With this in mind, the following list maps topics covered by week. I indicate our **actual class week** and the corresponding **module number** in the course room. We will chat more about this during the week 1 sync lecture (come with questions!).

Please note: In weeks 9 and 10 we will be covering 2 LMS modules, and in week 8 there is no corresponding LMS module.

PLEASE SUBMIT PROBLEM SETS BY ACTUAL WEEK and not LMS Module.

1. **Week 1**
    - **Module 1. Welcome to Data Science: Tools of the Trade**
        - **R Studio Crash Course and Debugging**
            - **Common Issues and Solutions**
            - **RMD Files**
            - **Linking Rstudio and Git – in class exercise.**
    - **Resources**
        - **Wickham: Introduction; Explore: Introduction; Workflow: Basics; Workflow: Projects Silver, Chapters 1–4**
            - **R Introduction and Resources Download R**
            - **R Basics**
            - **Download RStudio You want the Desktop version, free license**
        - **RStudio Introduction and Resources**
        - **There are many supporting resources found in the Repository**
            - **All items beginning with "01w_01m_"**
                - **(week 1 , module 1)**
    - **Lesson Notes**
        - **Chapter 1, Introduction**
        - **Synchronous Session: R basics, "verbs" of data wrangling**
    - **Assessments:**
        - **Problem Set 01w_01m due 24 hours before Week 3 live session**

2. **Week 2**
    - **Module 2. Analyzing Data: Conditional Means**
    - **Basic Coding and Debugging Exercises – in class exercise.**
    - **Knitting Basics – in class exercise.**
        - **Resources**
            - **Wickham: Data Transformation Silver, Chapters 5–9, 12–13 Lecture Notes**
            - **Chapter 2, Conditional Means**
            - **There are many supporting resources found in the Repository**
                - **All items beginning with "02w_02m_"**
                    - **(week 2 , module 2)**
        - **Synchronous Session: Conditional means**
    - **Assessments:**
        - **Problem Set 02w_02m due 24 hours before Week 3 live session**

3. **Week 3.**
    - **Module 4. Getting Data: Flat Files and "Tidy Data"**
        - **Resources**
            - **Wickham: Data Import; Tidy Data**
            - **There are many supporting resources found in the Repository**
                - **All items beginning with "03w_04m_"**
                    - **(week 3, module 4)**
        - **Async: Flat Data**

        - Synchronous Session: Working with various data formats
        - Assessments:
            - Problem Set 03w_04m due 24 hours before Week 4 live session

4. Week 4
    - Module 3. Presenting Data: Descriptive Plots … EDA!
        - Resources
            - Wickham: Data Visualization Data Transformation
            - Cookbook for R:Bar and Line Graphs Cookbook for R: Plotting Distributions
            - Chapter 3, Plotting Distributions and Conditional Means:
            - There are many supporting resources found in the Repository
                - All items beginning with "04w_03m_"
                    - (week 4 , module 3)
        - Synchronous Session: Presenting results in graphical format: bar plots, density plots, dot plots, histograms
        - Assessments:
            - Problem Set 04w_03m due 24 hours before Week 5 live session

5. Week 5.
    - Module 7. Getting Data: Web Sources + APIs
        - Resources
            - Rvest Vignette: https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html
            - There are many supporting resources found in the Repository
                - All items beginning with "05w_07m_"
                    - (week 5 , module 7)
        - Assessments:
            - Problem Set 05w_07m due 24 hours before Week 6 live session
        - Synchronous Session: Accessing data from the web
6. Week 6.
    - Module 11. Getting Data: Databases
    - Resources Wickham Relational Data
        - Working With Databases in R, available at https://dbplyr.tidyverse.org/articles/dbplyr.html
        - There are many supporting resources found in the Repository
            - All items beginning with "06w_11m_"
                - (week 6 , module 11)
    - Assessments:
        - Problem Set 06w_11m due 24 hours before Week 7 live session
        - Progress Report 2: Week 6 Project Report Deliverable
            - ** Due Prior to Week 6 Live Seminar. **
    - Lecture Notes
        - Chapter 11, Databases
    - Synchronous Session: Databases and relational data

7. **Week 7.**
    - **Module 12. Analyzing Data: Unsupervised Learning (k-means)**
        - **NEW CONTENT: HAC**
    - **Resources**
        - **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York, NY: Springer. Chapter 10 , Chapter 10 Lab R Code**
        - **There are many supporting resources found in the Repository**
            - **All items beginning with "06w_12m_"**
                - **(week 7 , module 12)**
    - **Assessments:**
        - **Problem Set 07w_12m due 24 hours before Week 8 live session**
    - **Lecture Notes**
        - **Chapter 12, Unsupervised Learning**
    - **Synchronous Session: K-means clustering, nearest neighbor classification**

8. **Week 8.**
    - **NEW CONTENT: Analyzing Data: ARM**
    - **Resources:**
        - **There are many supporting resources found in the Repository**
            - **All items beginning with "08w_xxm_"**
                - **(week 8 , module xx)**
                - **There is no supporting LMS Module**
    - **Assessments:**
        - **Problem Set 08w_xxm due 24 hours before Week 9 live session**

9. **Week 9.**
    - **Module 6. Presenting Data: Scatterplots**
    - **Module 5. Analyzing Data: Linear Regression**
    - **Resources**
        - **Wickham: Model: Introduction; Model Basics; Model Building**
        - **Wickham: Data Visualization, Graphics for Communication Tufte, Visual Display Chapters 4 and 5**
        - **Tufte, Envisioning Information, Chapter 2**
        - **There are many supporting resources found in the Repository**
            - **All items beginning with "09w_05m_" and "09w_06m_"**
                - **(week 9 , module 05) and (week 9 , module 06)**
    - **Lecture Notes**
        - **Chapter 5, Linear Regression**
        - **Chapter 6, Scatterplots**
    - **Synchronous Session: Using linear regression, training, and testing models, Presenting data via scatterplots**
    - **Assessments:**
        - **Problem Set 09w_56m due 24 hours before Week 9 live session**
        - **Progress Report 3: Week 9 Project Report Deliverable.**
            - **** Due Prior to Week 9 Live Seminar ****

10. Week 10.

- Module 8. Analyzing Data: Classification
  - NEW CONTENT: Decision Trees
- Module 9. Presenting Data: Plots and Tables for Classification
- Resources
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York, NY: Springer. Chapter 4 , Chapter 4 Lab R Code
  - Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014, May). How to ask for a favor: A case study on the success of altruistic requests. In ICWSM. Available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8106/8101
  - There are many supporting resources found in the Repository
    - All items beginning with "10w_09m_" and "10w_08m_"
      - (week 10 , module 9) and (week 10 , module 8)
- Lecture Notes
  - Chapter 8, Classification
  - Chapter 9, Plots and Tables for Classification
- Assessments:
  - Problem Set 10w_89m due 24 hours before Week 11 live session

11. Week 11.

- Module 10. Cross Validation
- Wickham: Many Models
- Lecture Notes
  - Chapter 10
- Resources:
  - There are many supporting resources found in the Repository
    - All items beginning with "11w_10m_"
      - (week 11 , module 10)
- Assessments:
  - Problem Set 11w_10m due 24 hours before Week 12 live session

12. Week 12.

- Module 14. Communicating Results (PRACTICE PRESENTATIONS)
- Assessments:
  - Progress Report 4: Week 12 Project Report Deliverable.
    - ** Due Prior to Week 12 Live Seminar **
- During Live Lecture: Practice Presentations

13. Week 13.

- Module 13. Presenting Data: Interactive Graphics (optional)
- NO Assessments Assigned
- Resources Lecture Notes

        - **Chapter 13, Interactive Graphics**
    - **Synchronous Session: Interactive graphics**
    - **During Live Lecture: Chat with Professor about Projects**

14. Week 14.

- **Module 14. Communicating Results (PRESENTATIONS)**
- **Assessments:**
    - **Week 14 Project FINAL REPORT. Due Prior to Week 14 Live Seminar.**

Final Project Milestones, Details and Rubric.

Deliverables.

- ~~Week 3 (Progress Report 1)~~
    - ~~o~~ We will skip PR 1. Everyone gets 100%
- Week 6 (Progress Report 2)
- Week 9 (Progress Report 3)
- Week 12 (Progress Report 4)
- Week 14 (Final Report)

Presentations.

- Week 12 (informal / practice)
    - Covers all topics covered in week 12 project deliverable
    - Presentations should be 7 minutes
    - Use many visualizations, plots, figures, tables, etc to concisely present findings.
- Week 14
    - Covers ALL topics covered in week 14 project deliverable
    - Presentations should be 7 minutes
    - Use many visualizations, plots, figures, tables, etc to concisely present findings.

Overview.

- You may work in teams of 2 - 4.
    - Please include ALL names on ALL submissions.
- You will be building a report and submitting your intermediate report periodically throughout the course.
- Week 12 and final report deliverables will be submitted as *BOTH* .RMD files and knitted counterpart (html, doc, or pdf).
- See details below. Read final project Rubric Document for Expectations on final project.

~~No Week 3 Deliverable (No Progress Report 1) !~~

Week 6 Deliverable (Progress Report 2). Due Prior to Week 6 Live Seminar.

- Goal: *Find optional team member* and *initially investigate* a Good Research Topic.
- Propose a "Data Science" Problem to solve using this data and subsequent analytics.
    - Clearly identify Problem Statement. Be sure Problem statement is clear.
    - If time, begin to search for data set.
- Write a brief paragraph summarizing your Data Science Problem.

Week 9 Deliverable (Progress Report 3). Due Prior to Week 9 Live Seminar.

- Goal: *Finalize your choice in* Dataset, and finalize your "Data Science" Problem to solve using this data and subsequent analytics. (Find a data set that interests you!)
    - Finding data sets:
        - Kaggle.com
        - https://toolbox.google.com/datasetsearch
- Write a 1 – 2 pages summarizing your Data Science Problem and Data Set.
    - Discuss (at least)
        - Data Acquisition
        - Data Format (file type, clean, …)
        - Number of observations
        - Number and type of variables
    - Use tables, figures, viz and examples of data.
    - Begin EDA if time.

Week 12 Deliverable (Progress Report 4). Due Prior to Week 12 Live Seminar.

- First Draft of report: 4-6 pages (not including Data Visualizations and Code)
- Sections
    - Introduction
        - Introduce problem.
        - Motivate your approach.
    - Data
        - How is data acquired?
        - Format of Data
        - Describe data / variables.
            - Quantitative, qualitative, etc.
        - If possible at this point, load data and create some supporting displays / visualizations. (This may not be possible yet if you plan to "scrape" your data.)
    - *EDA (Exploratory Data Analysis)*
        - *Investigate data: distribution of data, correlations, associations, and predictive potential to solve your proposed problem*
        - *Support investigation with excellent plots, charts, displays and visualizations.*
        - *REQUIREMENTS:*
            - *At least 3 different types of Visualizations*
    - *Models and Methods*
        - *Implement Classifiers, Models, Predictors, Clustering Results, ARM, etc to solve data science problem.*
        - *REQUIREMENTS:*
            - *Must include Clustering and/or ARM results*
            - *Must include Results from Regression Model AND Decision Tree Model.*
        - *Investigate the learned model and support with visualizations.*
        - *Report accuracy and reliability of results with relevant supporting viz.*

Final Product Deliverable. Due BEFORE Week 14 Live Seminar.

- FINAL Draft of report: 7-9 pages (not including Data Visualizations and Code)
- Sections
    - Introduction
        - Introduce problem.
        - Motivate your approach.
    - Data

        - How is data acquired?
        - Format of Data
        - Describe data / variables.
    - EDA (Exploratory Data Analysis)
        - Investigate data: distribution of data, correlations, associations, and predictive potential to solve your proposed problem
        - Support investigation with excellent plots, charts, displays and visualizations.
        - REQUIREMENTS:
            - At least 3 different types of Visualizations
    - Models and Methods
        - Implement Classifiers, Models, Predictors, Clustering Results, ARM, etc to solve data science problem.
        - REQUIREMENTS:
            - Must include Clustering and/or ARM results
            - Must include Results from Regression Model AND Decision Tree Model.
        - Investigate the learned model and support with visualizations.
        - Report accuracy and reliability of results with relevant supporting viz.
    - Experimental Design and Results.
        - Merge this into this new section: Report accuracy and reliability of initial results with relevant supporting viz.
        - Discuss the design of your experiment (crossvalidation). Explain why you chose the crossvalidation scheme chosen and indicate all factors that led to this decision (dataset size, distribution of classes, etc).
        - Present results and supporting viz. Include a Confusion Matrix, ROC curve, or appropriate visualization of results.
    - Concluding Remarks
        - Discuss conclusions of results and how they relate to the proposed problem.
        - Discuss Lessons Learned and Future Work.
    - References (as needed)
        - Cite data sources as appropriate
        - Feel free to use footnotes rather than reference section.

<u>Scoring for Final Project Final Deliverable</u>

Report

1. Technical Analysis	**/17
2. Graphical Presentation	**/17
3. Written Description	**/17
4. Organization, Clarity, Formatting	**/17
5. Coding	**/17

Presentation

1. Demonstration of understand of methods	**/5
2. Organized Slides with good Viz	**/5
3. Demonstration of understanding of analyses and conclusions	**/5

TOTAL **/100

Rubric:
Each portion, Out of 17 points.
16-17: Excellent!
14-15: Very good. Some details could be improved upon.
10-13: Notable concerns.
1. Technical Analysis **/17
2. Graphical Presentation **/17
3. Written Description **/17
4. Organization, Clarity, Formatting **/17
5. Coding **/17

Project EXEMPLARS:

*** Download before viewing ***

https://drive.google.com/file/d/1jOglRBwsMwp952ZjXOeBwT7o_cTShZ50/view?usp=sharing
https://drive.google.com/open?id=1U_b0eNtD3kkPZmw97vNvoI4kOvxeoi6q
https://drive.google.com/open?id=1OlPmkXeDMg4qyK-46LBYDIWdOPXMpnMQ

From RBloggers (great example of project from beginning to end):

https://www.r-bloggers.com/logistic-regression-in-r-a-classification-technique-to-predict-credit-card-default/