MATH 385
Introduction to Data Science

Section 01 Holt Hall 155 MoWeFr 1:00PM - 1:50PM

Edward A. Roualdes erouldes@csuchico.edu
Office Hours: Holt 204 MoWe 2-2:50, Community Coding in MLIB 442 TuTh 2-3:50, or by appointment

Resources

No formal textbook is required. We will instead make use of the following online books.

Boehmke, B. U of Cincinnati Buisiness Analysitcs R Programming Guide

Wickham, H. and Grolemund, G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. " O'Reilly Media, Inc.", 2016.

Further, the following resources will probably be incredibly helpful.

Broman, K. W. and Woo, K. H. 2018 Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

Rcpp: Seamless R and C++ Integration

Additional Requirements

- Access to a computer will be essential to master the material of this course.
- We will learn to code in R using the program RStudio.

Course Grading

Your final grade for this course will be given according to the $+/-$ grading system, based on the following percentages and scale: $90 - 100$, A; $80- < 90$, B; $70- < 80$, C; $60- < 70$, D; < 60, F.

Quizzes	10%
Participation	10%
Presentations	20%
Project	20%
Homework	40%

Quizzes

There will some quizzes to ensure everyone has at least covered the basics.

Participation

You are required to ask questions of other students and their presentations, at least sometimes.

Homework

All course homeworks are outlined below in Section Outline Homeworks. The integers enumerate nearly the minimum requirements. In fact, there is a list below this that details Extra Homework Requirements. You must complete at least the integer homeworks and 5 of the extra requirements. Students feeling the integer homeworks are too easy are encouraged to supplement the integer assignments with the fractional assignments.

For each homework you must turn something in and this thing must at least be an R Markdown document (and its source). If not an RMarkdown document, then you could turn in a Jupyter notebook, or slides that accompany your homework's presentation (see Section Presentations).

In general, your homeworks should include full sentences which explain the context of your work: what data set did you use, where did you get it (complete references), and what you will do with the data. This doesn't have to be long, but it does need to be complete enough for you to review the work a year from now and be able to explain it to a future employer.

Submit your homework to me directly or to my office (slide under door if I'm not there) no later than 5:00pm on the date the assignment is due. Working with other students on homework is allowed, subject to the Academic Integrity Policy below. After the due date, you are allowed to turn in homework before the next homework is due for up to 50% credit. After the next homework is due you will not be allowed to turn in late homework.

Presentations

You must present three of your homeworks and your Project. You can decide which homeworks you want to present. Come tell me when you want to present. Due to the size of our class, you can't all wait until the end of the semester to present. Since it is naturally easier to present simpler ideas, I will be grading harder for simpler presentation topics and grading easier for harder presentation topics.

For each presentation you must turn something in, most likely just your homework, and this thing must at least be an RMarkdown document (and its source). This could be the presentation itself, if there is code contained, but it doesn't have to be the presentation. Complete sentences are required in all written materials.

You are expected to present 4 times throughout the semester. Each homework presentation will be a maximum of 5 minutes. Time limits are strict. I will stop listening at the end of your time, not when you stop talking.

Students must either ask other students questions or comment, preferably positively, on the presentation their fellow students did. If you want to critique, use a critique sandwich: first a compliment, then a politly phrased critique, and then finish with a compliment. To do this well is harder than it sounds. Try to make the trio of comments flow. Finally, practice the critique sandwich, erring on the side of politeness.

Presntations will be graded out of 10 points. 10/10 points means you presented clearly, accurately, and answered everyone's (including my) questions adequately. 5/10 will be given for good faith attempts.

You can make up poor presentations by doubling up on a future presentation. For instance, suppose you received a poor grade on Simple Analysis. Later on, you might present on Simple Linear Regression Analysis. Say within this Simple Linear Regression Analysis presentation that you improve your effort on Simple Analysis. I will replace your Simple Analysis grade with your Simple Linear Regression Analysis grade, if it improves your Simple Analysis grade. Your Simple Linear Regression Analysis grade will be determined independently, since it involves new material.

You may not directly copy anybody's presentation, including your own previous presentations. At the very least, change the data set and then redo somebody's presentation.

I will absolutely not tolerate any judgement or rudeness to anybody about their presentation. This is to be a welcoming environment where we should not be afraid to present.

Further, don't be afraid to make a mistake publicly. Mistakes happen and in this course at least, you have multiple opportunities to learn from your mistakes without it hurting your grade.

Project

Project presentations will be a maximum of 20 mintues and a minimum of 10. Project presentations will take place during dead week and our scheduled final.

Your final project should be thought of as an extended presentation. You should attempt to present as many of the most complex homeworks as possible.

What you submit for your project must be appropriate to the medium of the project. If you write up some functions in an R package and decide to publish it on GitHub, then your package should have documentation and a README. If you make a tutorial using Bookdown, then you should have a short introduction, body, and conclusion, each of which consists of complete sentences.

The main differences between this and a homework presentation:

1. The expectation is for an analysis that reaches across as many of the topics covered in class as possible
2. A longer presentation
3. There are not make-ups
4. This should be something you could present to a future employer during an interview

Tests

There are no formal tests in this class.

Getting Help

- You can visit the Math Tutor Lab on the fourth floor of Meriam Library. You should also visit your instructor during his/her office hours.
- Free Tutoring by Appointment at the Student Learning Center.
- Help specific to R:
  - MATH 130 (#4793 or #5546) aka Math130
  - Chico R Users Group - RUG

  - Online, Free Introduction to R – Data Camp

- Me – though I reserve the right to refuse to provide help within 24 hours of an exam.

Diversity Policy

Respect: Students in this class are encouraged to speak up and participate during class meetings. Because the class will represent a diversity of individual beliefs, backgrounds, and experiences, every member of this class must show respect for every other member of this class.

Academic Integrity Policy

Students are permitted and encouraged to collaborate on all assignments other than examinations. However, each student must turn in their own work. Further, it is the expressed expectation of this instructor that all students demonstrate integrity and individual responsibility in all actions related to this course. Unethical behavior of any kind is unacceptable and will be prosecuted vigorously. Any sign of cheating in any way on any course exams or assignments will be addressed directly, according to university standards. If you do not understand what plagiarism is, or what cheating entails, you must seek information regarding this matter from the current University Catalog and from me. The consequences of plagiarism begin with a failing grade on the work, and possibly a failing grade in the course, depending upon university action. More information is found at http://catalog.csuchico.edu/viewer/15/STUDJUDAFFAIRS.html

Disability Support

If you have any disability related needs in terms of taking exams or other accommodations, please contact Disability Support Service (Colusa Hall 898-5959 or campus information 898-INFO for directions) on campus to obtain the appropriate documentation. Afterwards, come by my office and identify your needs within the first two weeks of class so that any necessary arrangements can be made.

Course Outline

- Introduction to R and RMarkdown
- Data format: CSV, importing/exporting data
- Tidy data
- Visualizations

- Exploratory data analysis
- Simple Linear Regression
- Functions
- Bootstrap and confidence intervals
- String manipulation
- Regular expressions
- Web scraping
- Data format: XML, converting to and from CSV
- Multiple Linear regression
- Logistic regression
- Clustering

Outline Homeworks

1. **Simple Analysis**
    - import data, tidy it if necessary
    - describe data
    - make a visualization of a variable or two
    - summarize a variable or two
    - interpret your summaries, in English and statistics

1.5 **Intermediate Analysis**

- Simple Analysis
- simple linear regression
- make a visualization of two variables at the same time
- calculate and interpret a bootstrap confidence interval about a mean and another statistic

2. **Intermediate Analysis**
    - Simple Analysis
    - simple linear regression

- make a visualization of three variables at the same time
- calculate and interpret a bootstrap confidence interval about a mean and another statistic

2.5 **Advanced Visualizations**

- Intermediate Analysis
- make a confidence interval plot, rug plot underneath a histogram, confidnece interval underneath a histogram, a violin plot with a confidence interval/box plot within, a bubble plot, a bubble plot on a map
- Must explain what your plot means
- If you choose this option, you must present as one of your presentation items

3. **Write a function of reasonable complexity**

    - RMarkdown slides
    - describe function's goal
    - walk through code
    - show function in action
    - use a default argument

3.5 **Write a function using Rcpp**

- RMarkdown slides
- describe function's goal
- walk through code
- show function in action
- use a default argument

4. **Scrape a Website**

    - use regex to scrape some data from a website
    - tidy the data
    - in the end you should have one data set

5. **Scrape a (possibly different) Website**

    - use regex to scrape some data from a website
    - tidy the data
    - in the end you should have one data set to analyze

- Intermediate Analysis

6. **Simple Linear Regression Analysis**

   - Simple Analysis
   - interpret an intercept and a slope of a simple linear model
   - interpret a confidence interval of a coefficient of a simple linear model
   - interpret a confidence interval of a prediction of a simple linear model
   - make a plot of simple linear regression

6.5 **Bootstrap Analysis**

   - Intermediate Analysis
   - bootstrap something other than a mean, median, or standard deviation
   - describe bootstrap
   - make an appropriate plot

7. **Multiple Linear Regression Analysis**

   - Intermediate Analysis
   - interpret an intercept and a slope of multiple regression
   - interpret a confidence interval of a coefficient of multiple regression
   - interpret a confidence interval of a prediction of multiple regression
   - interpret an intercept and a slope of logistic regression
   - interpret a confidence interval of a coefficient of logistic regression
   - interpret a confidence interval of a prediction of logistic regression

7.25 **Advanced Visualizations**

   - Multiple Linear Regression Analysis
   - Must explain what you're plot means
   - make a confidence interval plot, rug plot underneath a histogram, confidnece interval underneath a histogram, a violin plot with a confidence interval/box plot within, a bubble plot, a bubble plot on a map, OR a plot of multiple linear regression
   - If you choose this option, you must present as one of your presentation items

7.5 **Write Linear Regression Function**

   - start by writing a simple linear regression function in R – `linreg(x, y)`

- write a more complex model in R – `linreg(X, y)`
- coordinate descent
- demonstrate your function
- discuss why this was hard and/or what makes R's version better

7.75 **Regularization**

- Multiple Linear Regression Analysis
- fit a regularized multiple regression model – `glmnet::glmnet(x, y, alpha = 0); glmnet::cv.glmnet`
- show that regularization probably improves prediction accuracy, or if nothing else shrinks parameters – when might smaller parameters be beneficial?

8. **Logistic Regression Analysis**

- Intermediate Analysis
- interpret coefficients (plural) in model
- interpret prediction interval
- make a visualization of the model or the predictions
- draw a conclusion that might interest your boss

8.5 **Regularization**

- Logistic Regression Analysis
- fit a regularized logistic model
- show that regularization improves (technically at least doesn't hurt) prediction accuracy

9. **Clustering Analysis**

- Intermediate Analysis
- make an appropriate visualization
- describe why you chose this many groups
- make a visualization of the model or the predictions
- draw a conclusion that might interest your boss

Extra Homework Requirements
Must fulfill at least 5 of the following:

1. make RMarkdown slides

2. make RMarkdown notebook
3. make a Bookdown presentation
4. make Jupyter notebook
5. team presentation (counts as one presentation each), but all must actively participate in code and presentation
6. use LaTeX
7. use dplyr and/or tidyr
8. take on a more challenging version of any item
9. make an R package
10. use CIA factbook data
11. use geographic data
12. demonstrate/explain a previously unseen R code trick
13. write it in C++ via Rcpp

Project Ideas

1. **Set up Shiny Application**
    - Intermediate Analysis
    - interactive plots (plural)
    - interactive table
2. **Advanced Model Writing**
    - Intermediate Analysis
    - explain when, how, and why to use advanced model
    - model $\geq$ multiple linear regression
    - regularized
    - coordinate descent (probably)
    - demonstrate your model
    - Advanced Visualization
3. **Advanced++ Model Writing (any model)**

- Advanced Model Writing
- Write model using Rcpp
- Tutorial on model and Rcpp
- Advanced Visualization

4. **Advanced Bootstrap**
   - Intermediate Analysis
   - Bootstrap Analysis
   - bootstrap something other than a measure of location, spread, or correlation
   - Advanced Visualization
   - describe bootstrap in context of the data and the model
5. **Present a New Model**
   - Intermediate Analysis
   - Advanced Visualization
   - explain when to use it
   - explain how to use it
   - explain why it's better
   - demonstrate how it's an improvement over a previous model
   - Tutorial on new model with new data set
6. **Set up GitHub Repository**
   - repository and all code must be working
   - write README
   - use it from command line, not some gui
   - Tutorial on a model $\geq$ multiple linear regression or create an R package
7. **Set up (not GitHub hosted) R package**
   - Must fit a special model or plot
   - Can't just forward variables