DATA-LA 322

**Principles of Data Science**

Fall 2020

Tuesdays and Thursdays 12-1:15pm

Online

Lantz McGinnis-Brown

Office: Online

Office Hours: After class or by appointment

Email: [LantzBrown@boisestate.edu](mailto:LantzBrown@boisestate.edu)

Phone: (208) 386-0056

**Welcome!**

Welcome to DATA-LA 322! I look forward to getting to know each and every one of you! As your professor, I value your opinions and feedback on how to make this course better for you and others, so please feel free to communicate with me if there are ways I can make adjustments to allow for a more effective learning environment.

**Course Description:** Study of core concepts in data science in the liberal arts, including predictive modeling using machine learning and data mining; data gathering, extraction and cleaning; and exploratory data analysis. Course emphasizes practical skills for liberal arts students to examine questions of human behavior using large and complex data sets.

**Course Learning Outcomes:**

After successful completion of this course, students will be able to:

1. Utilize R statistical software to clean, organize, and visualize data
2. Analyze data using the appropriate statistical methods, including linear regression, logistic regression, and generalized linear models
3. Utilize tree-based methods for prediction
4. Execute cross validation to predict outcomes and understand issues of overtraining
5. Communicate data science processes and findings
6. Understand the ethical implications of data science as it is applied to human subjects and their data

**Course Format:**

This course will include lectures, in-class assignments, and a midterm and final exam (both take home).

**Prerequisites:**

This course requires CS133 (or CS111) and a statistics course as a prerequisite.  This previous experience gives you an overview of coding and an introduction to statistical thinking (probability theory, hypothesis testing). We will use that foundation to jump into the R statistical software.

**Texts (All freely available online):**

Grolemund, G. & Wickham, H. (2017) *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data.* Available at[**https://r4ds.had.co.nz/**](https://r4ds.had.co.nz/)

Irizarry, Rafael A. (2019) *Introduction to Data Science: Data Analysis and Prediction Algorithms with R.* Available at: <https://rafalab.github.io/dsbook/>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017) *An Introduction to Statistical Learning with Applications in R.* Available at <https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf>

**Software:**

R Statistical Software: <https://www.r-project.org/>

R Studio: <https://www.rstudio.com/products/rstudio/download/>

**Supplemental Resources:**

RStudio’s ‘cheatsheets’: <https://rstudio.com/resources/cheatsheets/>

**Course Website:**

We will be using Blackboard, a course management system which provides a website for me to post lecture materials, homework, readings, etc. All assignments will be posted to Blackboard and you can submit your assignments via Blackboard as well. You can find access to the site at: <https://blackboard.boisestate.edu>

**Course Requirements**: The course will consist of a mix of lectures by the instructor and in class activities. Grades will be based on a combination of: attendance/in-class activities, homework, and exams.

1. **Attendance & in-class activities**: You are required to attend class. This class covers a lot of material, and missing lectures can make it difficult to catch up. If you have an important reason to miss class, we can discuss ways for you to catch up on the material.
2. **Homework**: You will have homework assignments. Details will be provided in class. You are welcome to email me with specific questions on the homework, but I will not *check* your homework before it is due.

1. **Exams:** The two at-home examinations will consist of a data set that you need to examine. Content will be cumulative.

**Grading:**

Attendance: 20% \*One absence allowed without penalty

Homework: 40% (exceptions may be made under special

Exam 1: 20% circumstances)

Exam 2: 20% 100%

**Students needing accommodation:**

Students needing accommodations to fully participate in this class should contact the Educational Access Center (EAC). All accommodations must be approved through the EAC prior to being implemented.  To learn more about the accommodation process, visit the EAC’s website at <https://eac.boisestate.edu/new-eac-students/>.

**Student Conduct and Academic Integrity:**

In order to create a safe space for learning, I expect all of us to exhibit behavior that reflects Boise State’s Statement of Shared Values (<http://deanofstudents.boisestate.edu/statement-of-shared-values/)> and is characterized by

* Academic Excellence
* Caring
* Citizenship
* Fairness
* Respect
* Responsibility
* Trustworthiness

In addition, students in this course are expected to uphold standards outlined in the Boise State University Student Code of Conduct (<http://deanofstudents.boisestate.edu/student-code-of-conduct/>).

Any work submitted by a student in this course for academic credit will be the student's own work.

Note: While I encourage students to help each other on homework, if you submit the same document as another student, you will receive a 0%.

**Tentative Course Outline & Reading List:**

Please note that this list will be updated throughout the term depending on the amount of material we complete each class, and the topics that students are most interested in.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Date** | **Topic** | **Reading** | **Homework Due (will provide more details in class)** |
| Week 1 | 25-Aug | Introduction to Course & R Basics | [(W & G) Ch. 1 & 2](https://r4ds.had.co.nz/) | Download R & RStudio on personal computer |
| 27-Aug | R Basics Continued | (W & G) Ch. 27 | Intro R notebook |
| Week 2 | 1-Sep | Data Visualization (ggplot2) | (W & G) Ch. 3 | Basic Data Navigation HW |
| 3-Sep | Data Transformation (dplyr) & Pipes | (W & G) Ch. 5 & 18 | Visualization Homework |
| Week 3 | 8-Sep | Exploratory Data Analysis | (Irizzary) Ch. 8 | Data Transformation Homework |
| 10-Sep | Summarizing Data | (Irizzary) Ch. 11 |  |
| Week 4 | 15-Sep | Data Reshaping | (W & G) Ch. 12 | Data Exploration Homework |
| 17-Sep | Data Merging/Joining | (W & G) Ch. 13 |  |
| Week 5 | 22-Sep | Data Importing & Web Scraping | (Irizarry) Ch. 13 | Data Reshaping/Merging Homework |
| 24-Sep | Functional Programming | (W & G) Ch. 21 |  |
| Week 6 | 29-Sep | Intro to Modeling | (W & G) Ch. 23 | Functional Programming Homework |
| 1-Oct | Model Building | (W & G) Ch. 24 |  |
| Week 7 | 6-Oct | Many Models | (W & G) Ch. 25,  [Common statistical tests are linear models](https://lindeloev.github.io/tests-as-linear/) | Exploratory Linear Modeling HW |
| 8-Oct | Categorical Frequencies/Intro to Probability | (Irizarry) Ch. 13 |  |
| Week 8 | 13-Oct | Probability Distribution Functions | (Irizarry) Ch. 14 | Monte Carlo simulation HW |
| 15-Oct | Statistical Inference | (Irizarry) Ch. 15 |  |
| Week 9 | 20-Oct | Bayesian Models | (Irizarry) Ch. 16 | Probability Homework |
| 22-Oct | Intro to Machine Learning | (ISLR) Ch. 2 | Handout Exam #1 |
| Week 10 | 27-Oct | Regression | (ISLR) Ch. 3.1-3.3 | Work on Exam #1 |
| 29-Oct | Smoothing | (Irizarry) Ch. 28 | Exam #1 Due |
| Week 11 | 3-Nov | Classification | (ISLR) Ch. 4.1-4.5 | Regression HW |
| 5-Nov | Trees | (ISLR) Ch. 8.1 |  |
| Week 12 | 10-Nov | Evaluation (Accuracy vs Overfitting) | (ISLR) Ch. 2.2; [Google Machine Learning Crash Course - Classification Section](https://developers.google.com/machine-learning/crash-course/classification/) | Classification HW |
| 12-Nov | Evaluation (Cross-Validation & Bootstrapping) | (ISLR) Ch. 5 |  |
| Week 13 | 17-Nov | Ensembles (Boosting & Bagging) | (ISLR) Ch. 8.2 | Trees & Evaluation HW |
| 19-Nov | Unsupervised Learning (Clustering) | (ISLR) Ch. 10.3 |  |
| THANKSGIVING WEEK (Nov. 22-28) | | | |  |
| Week 14 | 1-Dec | Unsupervised Learning (Dimensionality Reduction) | (ISLR) Ch. 10.1-10.2 | Unsupervised Learning HW |
| 3-Dec | Machine Learning in Practice | (Irizarry) Ch. 32 |  |
| Week 15 | 8-Dec | Large Datasets (Data.Table) | (Irizarry) Ch. 33, [Data.Table Tutorial](https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html) | PCA/Machine Learning in Practice HW |
| 10-Dec | Review | Handout Final Exam | Data.Table HW |
| Final Exams | 17-Dec | Final Exam Due (11:59PM) | Final Exam Due |  |