STOR 320.01: Introduction to Data Science
Spring 2019

Instructor: Brendan Brown
E-mail: `bb@live.unc.edu`
Office: B26 Hanes Hall
Office hour: M 1:00-2:00 pm

Lectures: **MWF, 2:30 AM - 3:20 PM**, 120 Hanes Hall

Website: `brendanrbrown.github.io`

Assistant: Taylor Petty
E-mail: `tmpetty@live.unc.edu`
Office: B30 Hanes Hall
Office hours: T 2:00-3:00 pm, Th 4:45-5:45 pm

Description: This course is an application-driven introduction to data science. Statistical and computational tools are valuable from Silicon Valley startups, to marine biology labs, to news organizations. More importantly, they give us means to explore our world through data—even if you do not end up with a job in data science per se. These tools require technical skills such as programming and statistics. They also require professional skills such as communication, teamwork, problem solving, and critical thinking.
You will learn these tools and hone these skills through hands-on experience working with datasets provided in class and downloaded from certain public websites. More than half of the course is dedicated to fundamental technical skills in data science: the manipulation, preparation and exploration through summaries and graphs of data. We will cover additional topics to round out the skillset, such as modeling, web scraping and more advanced data presentation.
Come to every class with your computer and be ready to work with others. Using resources around you is a key component of successful data analysis.

Textbook: **R for data science**, by *Hadley Wickham*
available free online `https://r4ds.had.co.nz/`
and
An introduction to statistical learning, by *James et al*
available free online `https://faculty.marshall.usc.edu/gareth-james/ISL/`

Prerequisites: STOR 155 or an equivalent introductory statistics course.

Grade: Tutorials and participation (10 %)
Homework (35%)
Project (55%)

Contacting me:

- **If you email me by 5:00 pm on a given day, I will respond that day.**
- Otherwise, I will not respond until the following day.

Contacting TA: Will answer student emails some time between 1 and 5 pm, with some variability depending on the day.

Tutorials and participation:

The best way to learn coding is by imitation and repetition. Tutorials encourage you to imitate methods presented in class and will help build coding skills necessary for homework, the project and beyond.
There will be roughly one tutorial per class period, small tasks asking you to imitate methods presented in lecture. They will be split into one to three parts depending on length.
They are intended to be helpful, rather than to put pressure on you to perform. However, I do expect everyone to participate and will grade your submissions as described below.

- **Tutorials are intended to be completed without the instructors help.** Rely on internet searches, lecture materials, and your fellow students.
- You may work in groups of up to three people with students sitting adjacent to you, sharing code and markdown files.
- **However, each person** must produce their own html markdown file the completed tasks.
- I will select 10 people at random and ask their teams to email me the html files with their work, in class. Over the course of the semester, I will ensure each person has had a chance to submit tutorial work.
- As time allows, I will pick one of the submissions to review immediately in class.
- **Tutorials will be graded on effort and completion:** If you have demonstrated a valid attempt to complete the task correctly, you will receive full credit.

Tutorials will make up the totality of your participation grade.

However, I reserve the right to **penalize participation grades of students who are habitually disruptive in class**, e.g. habitual lateness, loud talking during lectures, and other behavior that degrades the experience of fellow students.

Homework:

- Homework is assigned weekly, in general.
- Homework is due by the beginning of class, on the given due date, by uploading the submission to Sakai.
- Each homework is worth the same amount toward your final grade.
- You may discuss homework with classmates and teaching staff. But you must submit your own work.
- You may and often should search online for solutions to coding problems. This is perfectly fine and encouraged.
- However, copying responses from students who have taken the course, including from sources online, is unacceptable and could be treated as an honor code violation.
- **Homework must be submitted as the html output from an R Markdown file.** In other words, your homework submission must be a .html file with all code and writing, as produced in R Markdown.

  Submissions that do not 'knit' to html will not be accepted. Such cases most often result from errors in the code, which students must correct before submission.
- Late homework submitted less than 24 hours from when it was due with have its **score reduced** 50%.
- Homework later than 24 hours or a failure to adhere to the rules above will result in a **score of zero for that assignment**.

Project: The final project is done in groups of at least 4 and worth a total of 100 points. There will be 4 parts of varying point values submitted throughout the semester. Groups will be randomly assigned.

- **Part one:** Project Proposal. Worth 10 points and will be due sometime in the middle of the semester after groups have been designated. Instructions and a due date will be released later and will give ample time to complete this part.
- **Part two:** Exploratory Data Analysis. Worth 20 points and will be due sometime towards the end of the semester after the Project Proposal has been completed.
- **Part three:** Final Paper. Worth 40 points and must be submitted on Sakai by 5:00PM on Friday, April 24.
- **Part four:** Final Presentation. Worth 30 points and will take place during the course's designated final exam time according to the university calendar. All materials for the presentation must be submitted on sakai by the **start of** the scheduled exam time.

Topics

Core programming and data science skills
– R Markdown
– data frame creation and manipulation
– summary statistics
– visualization
– exploratory data analysis
– 'tidy' and relational data
– functions and functional programming
– string manipulation and regular expressions
Modeling
– cross-validation
– classification techniques
– clustering
Advanced topics (time permitting)
– Shiny
– advanced modeling
– web scraping
– mapping

Honor Code: `http://instrument.unc.edu/`