# STOR 320-001: Introduction to Data Science
# Fall 2020

**Instructor:** Dr. Yao Li
E-mail: yaoli@email.unc.edu
Office Hours: Tuesday, Thursday 10:00AM to 11:00AM

**Assistant:** Kevin O Connor
E-mail: koconn@live.unc.edu
Office Hours:

Pavlos Zouboulоglou
E-mail: pavlos@live.unc.edu
Office Hours:

Sam Booth
E-mail: slbooth@live.unc.edu
Office Hours:

**Lectures:** Tuesday and Thursday 11:30AM – 12:45PM

**Labs:** 400 Friday 10:40AM – 11:30AM
401 Friday 12:00PM – 12:50PM
402 Friday 1:20PM – 2:10PM

**Course URL:** Website: https://liyao880.github.io/stor320/
Assignment Submission: https://sakai.unc.edu/ and login with your Onyen

**Zoom Links:** Due to the pandemic, lectures and office hours will be hosted live online via Zoom. Lectures will also be recorded and linked to course website.

Lectures:
Instructor Office Hours:
Kevin's Office Hours:
Pavlos's Office Hours:
Sam's Office Hours:

**Description:** This course is an application-driven introduction to data science. Statistical and computational tools are valued throughout the modern workplace from Silicon Valley startups, to marine biology labs, to Wall Street firms. 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. During the first part of the semester, we will focus on R programming skills and data visualization. Later topics will include: exploratory data analysis, data wrangling, modeling, and effective communication of results.

Plan to come to every class with your computer and ready to work with others. Using resources around you is a key component of successful data analysis. This includes the internet and people.

**Textbook:** **R for Data Science**, *by Hadley Wickham*.

available free online https://r4ds.had.co.nz/

**Prerequisites**: STOR 155 or an equivalent introductory statistics course.

**Final Grade:** Labs (30%)
Homework (30%)
Final Project (40%)

**Homework:** Homework are constructed using customized problems from real life data sets. These analyses allow you to practice the techniques learned from lab assignments.

- Each homework point is worth the same amount toward your final grade. So an assignment worth 80 points will be worth twice an assignment worth 40 points in 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 on Sakai. 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 will 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.

**Labs:** Labs are constructed using problems from the course textbook, *R for Data Science*. Each lab will be worth 20 points. These labs are to be completed using R Markdown and submitted as an HTML file on Sakai. Late submission submitted less than 24 hours from when it was due will have its score reduced 50%. Submission later than 24 hours or a failure to adhere to the rules above will result in a score of zero for that assignment.

**Final Project:** The final project is done in groups of **5** and worth a total of 100 points. There will be **4 parts** of varying point values submitted throughout the semester.

- Part I: **Project Proposal**, is worth **10 points** and will be due sometime in the middle of the semester after groups have been designated.

- Part II: **Exploratory Data Analysis**, is worth **20 points** and will be due sometime towards the end of the semester after the Project Proposal has been completed.
- Part III: **Final Paper**, is worth **40 points** and must be submitted on Sakai by **11:59PM** on **Friday, November 13**.
- Part IV: **Final Presentation**, is worth **30 points** and will take place during the last class, which is **11:00AM** on **Tuesday, November 17.** Slides must be submitted by **11:59PM** on **Monday, November 16**.

**Grade Scale:** Your final grade is based on a weighted average according to the previously addressed breakdown. Curving on individual/group assessments should not be expected. A curve may be applied to the final grades depending upon the class average. Conversion to a letter grade will be based on the table below:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 94 to 100 | B | 83 to 86.99 | C | 73 to 76.99 | D | 60 to 66.99 |
| A- | 90 to 93.99 | B- | 80 to 82.99 | C- | 70 to 72.99 | F | 0 to 59.99 |
| B+ | 87 to 89.99 | C+ | 77 to 79.99 | D+ | 67 to 69.99 | | |

These are hard break lines and no rounding will be applied to push an individual student up to a more desirable letter grade.

**Lectures:** 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
- linear and generalized linear models
- classification techniques
- clustering

Advanced topics

- Shiny
- more advanced modeling with support vector machines and tree-based methods
- web scraping

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