QTM 151: Introduction to Statistical computing II

Zhiyun Gong

Institute of Quantitative Theory and Methods

Emory University

Data analysis is a central and essential skill for any student in a field which utilizes the techniques of data, statistics, and quantitative methods. Imparting the necessary knowledge base and fine tuning of these skills is a crucial step in the education of undergraduate students in the quantitative sciences. To this end, it is my goal, in conjunction with the Piedmont Project, to bring Emory’s Sustainability Literacy Survey data to my statistical computing course. This particular course focuses on data analysis, and the teaching of statistical programming with real data will help bring the subject alive by giving students an opportunity to apply what they have learned to real world situations. Throughout the semester students will analyze the sustainability survey data. Along with the necessary theoretical background essential for understanding the big picture, the students will also experience a practical, hands-on approach to the analysis and interpretation of data. In particular, students will learn how to formulate their research questions, how to clean and transform data to useful and relevant information, how to model data to draw intelligent conclusions, and how to use these results to make appropriate recommendations (for future studies and to the scientific community in general). At the project presentation, students will communicate their findings. In addition to the analysis, there will be an emphasis on the sustainability aspect of these data driven project, and the students will be expected to explore that aspect of the project in a thorough and thoughtful manner.

**QTM 151: Introduction to Statistical Computing II**

Fall 2018

Zhiyun Gong, Ph. D.

## Contact Information

*Office*  Modern Languages Building, RM 408

*Office Phone* 404-727-4117

*E-Mail* zhiyun.gong@emory.edu

*Office hours* TBA

## Course Format

This is a lab session in statistical computing and taught by a regular faculty.

## Time/Locations

***Wednesday*** *(50 min)*

 **QTM 151** **11:00-11:50 am WH205**

## Required Materials

Textbook ***R for Data Science: Visualize, Model*, Transform, Tidy, and Import Data** by Hadley Wickham and Garrett Grolemund. This is just a strongly recommended textbook, not a requirement! There are a couple of other good books on statistical computing that I’d like to recommend to you as well:

* ***“ggplot2: Elegant Graphics for Data Analysis”,*** by Hadley Wickham (R package ggplot2)
* ***“A Handbook of Statistical Analyses using R” 3rd edition***, by T. Hothorn and B. Everitt

Computing All students are required to bring a laptop to labs. Labs use [R](https://www.r-project.org/) ([Rstudio](https://www.rstudio.com/)), a free statistical software. This is available for both Windows and Mac users, however older versions of Mac may require a system upgrade. Installation instructions are on Canvas.

## Course Description

This course provides a practicum of skills for data science and an introduction to how to do data science with R. The material is selected to enable you to get data into the most useful structure, transform it, visualize it, and model it. This will require you to practice the material outside of class.



**Learning Outcomes:**

By the end of the course, students should be able to (1) deal with complex and messy real data (2) use graphics to explore and understand data (3) gain familiarity with basic data manipulation, (4) fluently reshape data into the most convenient form for analysis, and (5) automate cleaning and analysis.

## Course Evaluation

Students’ performances will be evaluated through **attendance (5%),** **In-lab quizzes (35%),** **weekly homework (35%), and project (25%).** Grades are *not* rounded up at the end of the semester.

|  |  |
| --- | --- |
|  | Grading Scale |
|  | A | 93-100 | C+ | 77-79.9 | F | 0-59.9 |
|  | A- | 90-92.9 | C | 73-76.9 |   |   |
|  | B+ | 87-89.9 | C- | 70-72.9 |   |   |
|  | B | 83-86.9 | D+ | 67-69.9 |   |   |
|  | B- | 80-82.9 | D | 60-66.9 |   |   |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## Lab

Weekly lab material will generally be posted on Canvas before Wednesday class. The ratio of teaching to student practice is approximately 1:1. Additionally, content covered during labs may align or diverge from content in the textbook.

**Attendance (5%)**

Regular attendance is required.

## In-lab Exam (35%)

During the semester, there will be several in-lab quizzes, each student must complete the lab quizzes and submit their own work to Canvas during this lab session.

## Lab Homework (35%)

Lab homework is typically assigned every week and you will have approximately one week to complete each assignment. Lab homework may be created by the instructor. Unless prior arrangements are made well in advance, for reasons judged to be acceptable by me, late homework will not be accepted (as solutions will be posted soon after the due day). I’ll publish (anonymously) the best answer (corrected if necessary) each week as the solution so we can learn from each other. If you are not comfortable with this, please let me know.

Homework should be generated into Word by rmarkdown and submitted on Canvas. All work should be shown including R output, interpretation, and R script. You are encouraged to work with your group members but you have to submit your own work. The grader will grade lab homework. At the end of the semester, lab homework grades will be averaged to provide an overall lab homework score.

## Project (25%)

The data analysis project is designed to provide an overview of statistical data analysis. Students are required to write a one page written report and provide summary tables and graphics of data analysis. Students will be asked to give an in-class presentation on their project. There are multiple steps in submission process. More detailed instructions will be provided later in the semester.

|  |  |
| --- | --- |
| **Item** | **Tentative dates** |
| Project Proposal | Oct 12 |
| Preliminary submission | Nov 2 |
| Final presentation | Nov 30, Dec 7 |

## Dropped Grades

To allow for tough times and occasional forgetfulness, the lowest lab homework and the lowest lab quiz will be dropped at the end of the semester. Therefore, “I have been busy with another class,” is **not** a valid reason to request an extension.

## Canvas Discussion Forum

Questions related to the material covered in labs and homework should be posted on the Canvas Discussion Forum. When posting to the forum, please use a specific title for your post. Before posting a new question please make sure to check if your question has already been answered. I will be answering questions on the forum daily.

## Communication

* Questions regarding course content should be posted to the Canvas Discussion Forum.
* Email your instructor to notify them of lab absences and submit your late homework with your official documents.
* Ask questions during lab.
* Come to office hours to ask questions. Many programming ideas are much easier to explain in person rather than electronic communication.
* If you require a private appointment with me, please email me at least three available time slots you have (for ease of scheduling).

## Expectations

* All course materials and important announcements will be posted on Canvas. You are expected to check Canvas regularly and read your emails.
* You are responsible for checking your grades on Canvas to make sure everything is recorded correctly.
* You are responsible for notifying the instructor in a timely manner of any events that may adversely impact your performance in the class. You are responsible for discussing such events with the Office of Undergraduate Education.
* **You should expect to work 1 to 3 hours a week outside of lab on coursework.**  Some students may need to work more than that, and some students may need to work less.

Access and Disability ResourcesStudents with medical/health conditions that might impact academic success should visit Access, Disability Services and Resources (ADSR) to determine eligibility for appropriate accommodations. Students who receive accommodations must present the Accommodation Letter from ADSR to the instructor at the beginning of the semester, or when an accommodation is granted.

## Academic Integrity

Upon every individual who is a part of Emory University falls the responsibility for maintaining in the life of Emory a standard of unimpeachable honor in all academic work. The [Honor Code of Emory College](http://catalog.college.emory.edu/academic/policy/honor_code.html) is based on the fundamental assumption that every loyal person of the University not only will conduct his or her own life according to the dictates of the highest honor, but will also refuse to tolerate in others action which would sully the good name of the institution. Academic misconduct is an offense generally defined as any action or inaction which is offensive to the integrity and honesty of the members of the academic community.  **The typical sanction for a violation of the Emory Honor Code is an F in the course.**  **Any suspected case of academic misconduct will be referred to the Emory Honor Council.**

## Tentative Schedule (Fall 2018)

This is a tentative outline of the schedule and is subject to change.

|  |  |
| --- | --- |
| **Date (Fridays)** | **Topic Covered** |
| Aug 31 | Introduction to QTM 151 and Data import |
| Sep 7 | Data visualization: package ‘ggplot2’ |
| Sep 14 | Data visualization: package ‘plotly’  |
| Sep 21 | Data transformation: package ‘dplyr’ |
| Sep 28 | Data transformation: package ‘dplyr’ |
| Oct 5 | Data tidy-up: package ‘tidyr’ and intro to project  |
| Oct 12 | Data tidy-up: package ‘tidyr’  |
| Oct 19 | Data manipulation: package ‘stringr’ |
| Oct 26 | Text mining  |
| Nov 2 | Data functional package: ‘purrr’ |
| Nov 9 | Basic models and linear regression |
| Nov 16 | Interactive Web: shiny |
| Nov 30 | Github |
| Dec 7 | Project Presentation  |
| Apr 27 | Project Presentation |