Boston University Metropolitan College

Data Analysis and Visualization with R

MET CS555 A3

Guanglan Zhang
guanglan@bu.edu
Office hours: Thursday 2-4pm
Classroom Location: MET, 1010 Commonwealth Avenue, Room 101

Course Description

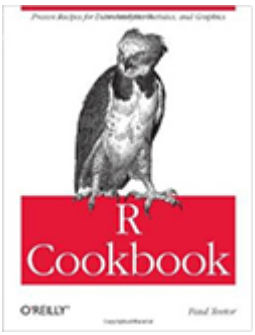
This course provides an overview of the statistical tools most commonly used to process, analyze, and visualize data. Topics include describing data, statistical inference, 1 and 2 sample tests of means and proportions, simple linear regression, multiple regression, logistic regression, analysis of variance, and regression diagnostics. These topics are explored using the statistical package R, with a focus on understanding how to use and interpret output from this software as well as how to visualize results. In each topic area, the methodology, including underlying assumptions and the mechanics of how it all works along with appropriate interpretation of the results, are discussed. Concepts are presented in context of real world examples.

Prerequisites

CS546 (Quantitative Methods for Information Systems) and CS544 (Foundations of Analytics) or equivalent background.

Required Book

The following two books are required for the course. This should be used as a reference to help support you in your assignments and supplementing the course's Classroom sessions on R.

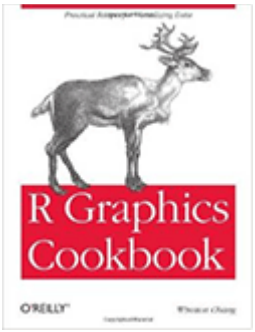


Long, J. D., & Teetor, P. (2019). R cookbook: proven recipes for data analysis, statistics, and graphics, CA: O'Reilly.

It is freely available online at: https://rc2e.com/.

Recommended Book

Boston University Metropolitan College



Chang, W. (2018). R graphics cookbook: practical recipes for visualizing data. O'Reilly Media.

Courseware
There are six online modules covering the course content in the Blackboard site.

BU Community COVID-19 Public Health Policies

All students returning to campus will be required to be vaccinated against COVID-19, and upload information about their status (including applications for a medical or religious exemption or an extension) to the Patient Connect portal. In addition to the vaccine requirement, students must follow all other safety protocols, including the face covering policy, and screening, contact tracing, and testing requirements. At the beginning of each class you will be asked to show a green Healthway compliance badge on your mobile device to the instructor, and wear your face mask over your mouth and nose at all times.

Class Policies

1) **Attendance & Absences** – Full attendance and participation is expected. If there is a reason to miss a session, advanced notice through email should be sent to the lecturer.
2) **Assignment Completion & Late Work** – All assignments should be submitted on time. If there is a delay, the student must be in touch with the instructor. Late submissions without reasons will result in grade deduction.
3) **Academic Conduct Code** –Cheating and plagiarism will not be tolerated in any Metropolitan College course. They will result in no credit for the assignment or examination and may lead to disciplinary actions. Please take the time to review the Student Academic Conduct Code: http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/code.html.

   NOTE: [This should not be understood as a discouragement for discussing the material or your particular approach to a problem with other students in the class. On the contrary – you should share your thoughts, questions and solutions. Naturally, if you choose to work in a group, you will be expected to come up with more than one and highly original solutions rather than the same mistakes.]

Grading Criteria

- Homework Assignments and Term Project
  The six homework assignments are focused on applying theory learned in the week's module to a set of data and analyzing that data in R. Assignment submissions should be a single Microsoft Word or PDF file. The R code used to generate your results should be appended to the end of your assignment. Term project at the end of the semester gives you freedom to select a research question of your interest and answer it by applying what you have learnt in the course.

- Quizzes
  The six quizzes will evaluate students understanding of concepts presented in the corresponding week's module. Students should ensure adequate preparation before starting the quiz. It will not be possible to do well on the quiz without first reviewing the course material in depth and attempting to understand all examples and test yourself questions. It is recommended that you complete the quiz after you feel comfortable with the material and asked any questions that you may have had.

- Midterm Examination and Final Examination
  The midterm exam will cover material from module 1-3. The final exam will be comprehensive and will cover material from the entire course. Both will be open-book exams.

The final grade for this course will be based on the following:

Deliverable	**Weight**
Assignments	20%
Quizzes	15%
Midterm Exam	20%
Class Participation	5%
Term Project	10%
Final Exam	30%

Class Meetings, Lectures & Assignments

Lectures, Readings, and Assignments subject to change, and will be announced in class as applicable within a reasonable time frame.

Date	Topic	Readings Due	Due dates
Sept. 2 Lecture 1	• Fundamental Elements of Statistics • Qualitative and Quantitative Data Summaries	Online Module 1 Textbook sections 2.6, 2.13, 3.1, 3.6, 3.9	
Sept. 9	• Normal distribution	Online Module 1	

Lecture 2	• Sampling • The Central Limit Theorem	Textbook sections 8.9, 8.10, 8.11, 9.1, 9.2, 9.5, 10.9, 10.11, 10.16, 10.18	
Sept. 16 Lecture 3	• Statistical Inference • Confidence Intervals • Test of Significance • Stating Hypotheses • Test Statistics and p-Values • Evaluating Hypotheses	Online Module 2 Textbook sections chapter 9 introduction, 9.8, 9.9, 9.15	Assignment 1 Quiz 1
Sept. 23 Lecture 4	• Significance Test "Recipe" • Significance Tests and Confidence Intervals • Inference about a Population Mean • Two-Sample Problems	Online Module 2 Textbook sections 10.10, 10.17	
Sept. 30 Lecture 5	• Scatterplots • Correlation	Online Module 3 Textbook sections 9.17, 10.1, 10.6	Assignment 2 Quiz 2
Oct. 7 Lecture 6	• Simple Linear Regression • F-test for Simple Linear Regression • t-test for Simple Linear Regression	Online Module 3 Textbook sections chapter 11 introduction, 11.1, 11.3, 11.4, 11.13	
Oct. 14 Lecture 7	• Residual Plots • Outliers and Influence Points • Assumptions of least-square regression	Online Module 4	Assignment 3 Quiz 3
Oct. 21	Midterm Exam		
Oct. 28 Lecture 8	• Equation of multiple linear regression • Interpretation of multiple linear regression • F-test for Multiple Linear Regression • t-tests in Multiple Linear Regression • Cautions about Regression	Online Module 4 Textbook sections 11.2, 11.8, 11.10, 11.11, 11.14	
Nov. 4 Lecture 9	• One-Way Analysis of Variance • F-test for ANOVA • Evaluating Group Differences • Type I and Type II Errors	Online Module 5 Textbook sections 11.20, 11.21, 11.22	Assignment 4 Quiz 4
Nov. 18	• Issues with Multiple Comparisons	Online Module 5	

Lecture 10	• Assumptions of Analysis of Variance • Relationship between One-Way Analysis of Variance and Regression • One-Way Analysis of Covariance • Two-Way Analysis of Variance • Two-Way Analysis of Covariance		
Nov. 18 Lecture 11	• One-Sample Tests for Proportions • Significance Tests for a Proportion • Confidence Intervals for a Proportion	Online Module 6 Textbook sections 9.11, 9.12,	Assignment 5 Quiz 5
Dec. 2 Lecture 12	• Two-Sample Tests for Proportions • Confidence Intervals for Differences in Proportions • Significance Tests for Differences in Proportions • Effect Measures • Logistic Regression • Multiple Logistic Regression • Area under the ROC Curve	Online Module 6 Textbook sections 9.18, 13.7	
Dec. 9 Lecture 13	Review session		Assignment 6 Quiz 6 Term Project
Dec. 16	Final Exam		

Instructor Biography

Guanglan Zhang, Ph.D.

Dr. Guanglan Zhang received her Ph.D. from School of Computer Engineering, Nanyang Technological University, Singapore for doctoral work in bioinformatics. She is an Associate Professor in Computer Science at Boston University Metropolitan College. Dr. Zhang has worked in the data mining and data analytics field since 1998. The most important aspects of her work include biomedical data analysis, development and implementation of biomedical databases, computational simulations of laboratory experiments, development of diagnostic methods for tissue typing, and computational support for vaccine development. Computational tools that she developed are used in the study of immunology, vaccinology, infectious disease, and cancer. She has authored more than 50 peer-reviewed scientific journal publications and developed dozens of biomedical and computational systems.