Boston University Metropolitan College

Data Analysis and Visualization

CS555 D1

Guanglan Zhang
guanglan@bu.edu
Office hours: Thursday 2-5pm at 808 Commonwealth Avenue, Room 250
Office Location: CAS, 685 - 725 Commonwealth Avenue, Room 324, Boston
Phone: (617) 358-5688

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.

Learning Objectives

By successfully completing this course you will be able to:

- Appreciate the science of statistics and the scope of its potential applications
- Summarize and present data in meaningful ways
- Select the appropriate statistical analysis depending on the research question at hand
- Form testable hypotheses that can be evaluated using common statistical analyses
- Understand and verify the underlying assumptions of a particular analysis
- Effectively and clearly communicate results from analyses performed to others
- Conduct, present, and interpret common statistical analyses using R

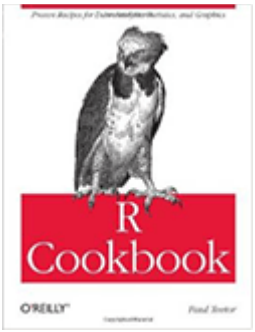
Prerequisites

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

Required Book

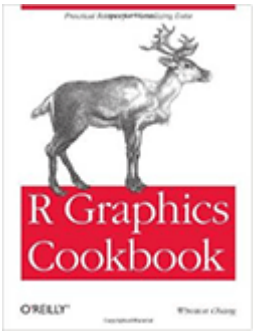
The following two books are required for the course. This should be used as a reference to help support you in your assignments and supplementing the course's Live Classroom sessions on R.

Boston University Metropolitan College



Teetor, P. (2011). *R cookbook.* Sebastopol, CA: O'Reilly. ISBN 9780596809157.

Recommended Book



Chang, W. (2013). *R graphics cookbook.* Sebastopol, CA: O'Reilly. ISBN 9781449316952.

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
  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. Lectures relating to R will be held and recordings will be posted after each session. Slides from the session will also be made available to students.

- 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.

- Final Examination
  The final exam will be comprehensive and will cover material from the entire course. It will be an open-book proctored exam consisting of questions similar to the ones in the quizzes but longer in length.

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

Deliverable	**Weight**
Weekly Homework Assignments	30%
Weekly Quizzes	30%
Classroom Participation	5%
Final Exam	35%

Study Guide

Lecture 1 Introduction to the science of statistics part 1

- Fundamental Elements of Statistics
- Qualitative and Quantitative Data Summaries

Lecture 2 Introduction to the science of statistics part 2

- Normal distribution
- Sampling
- The Central Limit Theorem

Lecture 3 Confidence intervals and hypothesis tests part 1

- Statistical Inference
- Confidence Intervals
- Test of Significance
- Stating Hypotheses
- Test Statistics and p-Values
- Evaluating Hypotheses

Lecture 4 Confidence intervals and hypothesis tests part 2

- Significance Test "Recipe"
- Significance Tests and Confidence Intervals
- Inference about a Population Mean
- Two-Sample Problems

Lecture 5 Understanding the association between two continuous or quantitative factors part 1

- Scatterplots
- Correlation

Lecture 6 Understanding the association between two continuous or quantitative factors part 2

- Simple Linear Regression
- F-test for Simple Linear Regression
- t-test for Simple Linear Regression

Lecture 7 Regression diagnostics

- Residual Plots
- Outliers and Influence Points
- Assumptions of least-square regression

Lecture 8 Multiple linear regression

- 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

Lecture 9 Analysis of Variance (ANOVA) part 1

- One-Way Analysis of Variance
- F-test for ANOVA
- Evaluating Group Differences
- Type I and Type II Errors

Lecture 10 Analysis of Variance (ANOVA) part 2

- Issues with Multiple Comparisons
- 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

Lecture 11 Analysis for proportions part 1

- One-Sample Tests for Proportions
- Significance Tests for a Proportion
- Confidence Intervals for a Proportion

Lecture 12 Analysis for proportions part 2

- 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

Lecture 13 Review session

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 Assistant Professor in Computer Science at Boston University Metropolitan College. She is also holding an adjunct position at Dana-Farber Cancer Institute and Harvard Medical School.

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 40 peer-reviewed scientific journal publications and developed dozens of biomedical and computational systems.

Boston University Library Link

As Boston University students you have full access to the BU Library—even if you do not live in Boston. From any computer, you can gain access to anything at the library that is electronically formatted. To connect to the library use the link http://www.bu.edu/library. You may use the library's content whether you are connected through your online course or not, by confirming your status as a BU community member using your Kerberos password.

Once in the library system, you can use the links under "Resources" and "Collections" to find databases, eJournals, and eBooks, as well as search the library by subject. Go to http://www.bu.edu/library/research/collections to access eBooks and eJournals directly. If you have questions about library resources, go to http://www.bu.edu/library/help/ask-a-librarian to email the library or use the live chat feature.

To locate course eReserves, go to http://www.bu.edu/library/services/reserves.

Please note that you are not to post attachments of the required or other readings in the water cooler or other areas of the course, as it is an infringement on copyright laws and department policy. All students have access to the library system and will need to develop research skills that include how to find articles through library systems and databases.

Academic Conduct Policy

For the full text of the academic conduct code, please go to http://www.bu.edu/met/for-students/met-policies-procedures-resources/academic-conduct-code/.

A Definition of Plagiarism

"The academic counterpart of the bank embezzler and of the manufacturer who mislabels products is the plagiarist: the student or scholar who leads readers to believe that what they are reading is the original work of the writer when it is not. If it could be assumed that the distinction between plagiarism and honest use of sources is perfectly clear in everyone's mind, there would be no need for the explanation that follows; merely the warning with which this definition concludes would be enough. But it is apparent that sometimes people of goodwill draw the suspicion of guilt upon themselves (and, indeed, are guilty) simply because they are not aware of the illegitimacy of certain

kinds of "borrowing" and of the procedures for correct identification of materials other than those gained through independent research and reflection."

"The spectrum is a wide one. At one end there is a word-for-word copying of another's writing without enclosing the copied passage in quotation marks and identifying it in a footnote, both of which are necessary. (This includes, of course, the copying of all or any part of another student's paper.) It hardly seems possible that anyone of college age or more could do that without clear intent to deceive. At the other end there is the almost casual slipping in of a particularly apt term which one has come across in reading and which so aptly expresses one's opinion that one is tempted to make it personal property."

"Between these poles there are degrees and degrees, but they may be roughly placed in two groups. Close to outright and blatant deceit-but more the result, perhaps, of laziness than of bad intent-is the patching together of random jottings made in the course of reading, generally without careful identification of their source, and then woven into the text, so that the result is a mosaic of other people's ideas and words, the writer's sole contribution being the cement to hold the pieces together. Indicative of more effort and, for that reason, somewhat closer to honest, though still dishonest, is the paraphrase, and abbreviated (and often skillfully prepared) restatement of someone else's analysis or conclusion, without acknowledgment that another person's text has been the basis for the recapitulation."

The paragraphs above are from H. Martin and R. Ohmann, *The Logic and Rhetoric of Exposition, Revised Edition.* Copyright 1963, Holt, Rinehart and Winston.

Academic Conduct Code

I. **Philosophy of Discipline**

The objective of Boston University in enforcing academic rules is to promote a community atmosphere in which learning can best take place. Such an atmosphere can be maintained only so long as every student believes that his or her academic competence is being judged fairly and that he or she will not be put at a disadvantage because of someone else's dishonesty. Penalties should be carefully determined so as to be no more and no less than required to maintain the desired atmosphere. In defining violations of this code, the intent is to protect the integrity of the educational process.

II. Academic Misconduct

Academic misconduct is conduct by which a student misrepresents his or her academic accomplishments, or impedes other students' opportunities of being judged fairly for their academic work. Knowingly allowing others to represent your work as their own is as serious an offense as submitting another's work as your own.

III. Violations of this Code

Violations of this code comprise attempts to be dishonest or deceptive in the performance of academic work in or out of the classroom, alterations of academic records, alterations of official data on paper or electronic resumes, or unauthorized collaboration with another student or students. Violations include, but are not limited to:

A. **Cheating on examination**. Any attempt by a student to alter his or her performance on an examination in violation of that examination's stated or commonly understood ground rules.

B. **Plagiarism.** Representing the work of another as one's own. Plagiarism includes but is not limited to the following: copying the answers of another student on an examination, copying or restating the work or ideas of another person or persons in any oral or written work (printed or electronic) without citing the appropriate source,

and collaborating with someone else in an academic endeavor without acknowledging his or her contribution. Plagiarism can consist of acts of commission-appropriating the words or ideas of another-or omission failing to acknowledge/document/credit the source or creator of words or ideas (see below for a detailed definition of plagiarism). It also includes colluding with someone else in an academic endeavor without acknowledging his or her contribution, using audio or video footage that comes from another source (including work done by another student) without permission and acknowledgement of that source.

C. **Misrepresentation or falsification of data** presented for surveys, experiments, reports, etc., which includes but is not limited to: citing authors that do not exist; citing interviews that never took place, or field work that was not completed.

D. **Theft of an examination**. Stealing or otherwise discovering and/or making known to others the contents of an examination that has not yet been administered.

E. **Unauthorized communication during examinations**. Any unauthorized communication may be considered prima facie evidence of cheating.

F. **Knowingly allowing another student to represent your work as his or her own**. This includes providing a copy of your paper or laboratory report to another student without the explicit permission of the instructor(s).

G. **Forgery, alteration, or knowing misuse of graded examinations, quizzes, grade lists, or official records of documents**, including but not limited to transcripts from any institution, letters of recommendation, degree certificates, examinations, quizzes, or other work after submission.

H. **Theft or destruction of examinations or papers** after submission.

I. **Submitting the same work in more than one course** without the consent of instructors.

J. **Altering or destroying another student's work or records**, altering records of any kind, removing materials from libraries or offices without consent, or in any way interfering with the work of others so as to impede their academic performance.

K. **Violation of the rules governing teamwork**. Unless the instructor of a course otherwise specifically provides instructions to the contrary, the following rules apply to teamwork: 1. No team member shall intentionally restrict or inhibit another team member's access to team meetings, team work-in-progress, or other team activities without the express authorization of the instructor. 2. All team members shall be held responsible for the content of all teamwork submitted for evaluation as if each team member had individually submitted the entire work product of their team as their own work.

L. **Failure to sit in a specifically assigned seat during examinations.**

M. **Conduct in a professional field assignment that violates the policies and regulations of the host school or agency.**

N. **Conduct in violation of public law occurring outside the University that directly affects the academic and professional status of the student, after civil authorities have imposed sanctions.**

O. **Attempting improperly to influence the award of any credit, grade, or honor.**

P. **Intentionally making false statements to the Academic Conduct Committee or intentionally presenting false information to the Committee.**

Q. **Failure to comply with the sanctions imposed under the authority of this code.**

Disability Services

Boston University makes every effort to accommodate the unique needs of its students. In keeping with university policy, students are expected to contact the Office of Disability Services (ODS) (www.bu.edu/disability/) each time they register for a course to request accommodations for that course.

Any student who feels he or she may need an accommodation for a documented disability should contact the Office for Disability Services at (617) 353-3658 or at access@bu.edu for review and approval of accommodation requests.

Technical Support

Experiencing issues with BU websites or Blackboard?

It may be a system-wide problem. Check the BU Information Services & Technology (IS&T) news page for announcements.

Boston University technical support via email (ithelp@bu.edu), the support form (http://www.bu.edu/help/tech/), and phone (888-243-4596) is available from 8 AM to midnight eastern time. For other times, you may still submit a support request via email, phone, or the support form, but your question won't receive a response until the following day. If you aren't calling, it is highly recommended that you submit your support request via the technical-support form at http://www.bu.edu/help/tech/learn as this provides the IS&T Help Center with the best information in order to resolve your issue as quickly as possible.

Examples of issues you might want to request support for include the following:

- Problems viewing or listening to sound or video files
- Problems accessing internal messages
- Problems viewing or posting comments
- Problems attaching or uploading files for assignments or discussions
- Problems accessing or submitting an assessment

To ensure the fastest possible response, please fill out the online form using the link below:

IT Help Center Support	
Web	http://www.bu.edu/help/tech/learn
Phone	888-243-4596 or local 617-353-4357
Check your open tickets using BU's ticketing system.	