Boston University Metropolitan College

Data Analysis and Visualization

MET - CS 555

Farshid Alizadeh-Shabdiz, PhD, MBA
alizadeh@bu.edu
Office hours: by appointment

Course Description

This course provides an overview of the most commonly used statistical tools to process, analyze, and visualize a data set. Topics include how to describe data, statistical inference, 1 and 2 sample tests of means and proportions, simple linear regression, multiple linear regression, multinomial 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 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

Text Books

Boston University Metropolitan College

The following two books are required for the course. These should be used as reference material to help support you in your assignments and supplementing the course's Live Classroom sessions on R. The modules themselves will provide you with the necessary information for the theory, concepts, and examples that you will need to complete your quizzes and understand the methodologies that you will apply to the problems presented in the homework assignments.

There will be no reading assignments from these books. These are excellent supplemental texts that you may want to review as we go through the course and also keep as reference text as you continue to use R in the future.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) An Introduction to Statistical Learning with Applications in R. Springer.
  The book has been make available online at https://statlearning.com

- Teetor, P. (2019). **R cookbook.** Sebastopol, CA: O'Reilly. ISBN -13: 978-1492040682
  The book has been made available online at https://rc2e.com and
  the code at https://github.com/CerebralMastication/R-Cookbook

- Chang, W. (2021). R graphics cookbook. Sebastopol, CA: O'Reilly. ISBN 9781491978573
  The book has been made available online at https://r-graphics.org
  And the code at https://github.com/wch/rgcookbook

Additional Text Books for further reading:

- **Andy Field, Jeremy Miles and Zoe Field. (2012) Discovering Statistics Using R.**
  Publisher: SAGE Publications Ltd. ISBN-13: 978-1446200469

- https://www.openintro.org/stat/ Free PDF for download & R tutorials and codes.

Additional Reference Books

- "Using R for Introductory Statistics, 2nd edition", by John Verzani, CRC Press, 2014. ISBN13: 978- 1466590731. (Reference book)

- "R for Everyone: Advanced Analytics and Graphics, 2nd Edition", by Jared P. Lander, Addison-Wesley Professional, 2017. ISBN13: 978-0134546926. (Reference book)

Courseware
List course website (Blackboard), as well as any web links that will be necessary for the class.

COVID-19 Policies

Compliance: All students returning to campus will be required, through a digital agreement, to commit to a set of Health Commitments and Expectations including face coverings, symptom attestation, testing, contact tracing, quarantine, and isolation. The agreement makes clear that compliance is a condition of being a member of our on-campus community.

You have a critical role to play in minimizing transmission of COVID-19 within the University community, so the University is requiring that you make your own health and safety commitments. Additionally, if you will be attending this class in person, you will be asked to show your Healthway badge on your mobile device to the instructor in the classroom prior to starting class, and wear your face mask over your mouth and nose at all times. If you do not comply with these rules you will be asked to leave the classroom. If you refuse to leave the class, the instructor will inform the class that they will not proceed with instruction until you leave the room. If you still refuse to leave the room, the instructor will dismiss the class and will contact the academic Dean's office for follow up.

Boston University is committed to offering the best learning environment for you, but to succeed, we need your help. We all must be responsible and respectful. If you do not want to follow these guidelines, you must participate in class remotely, so that you do not put your classmates or others at undue risk. We are counting on all members of our community to be courteous and collegial, whether they are with classmates and colleagues on campus, in the classroom, or engaging with us remotely, as we work together this fall semester.

Class Policies

1) **Assignment Completion & Late Work** – all the assignment has to be submitted in person or electronically (e.g. email). No late work will be acceptable.
2) **Laptop Requirement –** Students should have a personal laptop. We will use laptops in class room to write R program code. Also for the final exam you will need a Laptop.

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.

Grading Criteria

The course grade will be based on

- Active class participation (10%)
- Quizzes (20%)
- Assignments (20%)
- Final project (25%)
- and final exam (25%)

Assignments are expected to be submitted by their respective due dates. Late submissions are not accepted.

Homework Assignments

There will be homework assignments focused on applying theory learned in the class to analyze a data set in R. Assignment submissions should be in a single **Microsoft Word or PDF** file. The R code used to generate your results should be appended to the end of your assignment.

Quizzes

There will be six quizzes to assess students understanding of concepts presented in the class. Students should ensure adequate preparation before starting the quiz. Please note that it won't be possible to do well on the quizzes without reviewing the course material in.

Final Examination

The final exam will be comprehensive and will cover material from the entire course.

Study Guide

Introduction to the science of statistics

- Fundamental Elements of Statistics
- Qualitative and Quantitative Data Summaries
- Normal distribution
- Sampling
- The Central Limit Theorem

Confidence intervals and hypothesis tests

- Statistical Inference
- Stating Hypotheses
- Test Statistics and p-Values
- Evaluating Hypotheses
- Significance Test “Recipe”
- Significance Tests and Confidence Intervals
- Inference about a Population Mean
- Two-Sample Problems

Understanding the association between two continuous or quantitative factors

- Scatterplots
- Correlation

Linear Regression

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

Regression diagnostics

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

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

Analysis of Variance (ANOVA)

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

Analysis for proportions

- One-Sample Tests for Proportions
- Significance Tests for a Proportion
- Confidence Intervals for a Proportion
- Two-Sample Tests for Proportions
- Confidence Intervals for Differences in Proportions
- Significance Tests for Differences in Proportions
- Effect Measures

Logistic Regression

- Logistic Regression
- Multiple Logistic Regression