**Boston University** Metropolitan College

# Foundations of Machine Learning
MET CS555 A3 SPRG23 (Thursdays, 6PM)

**Bld CAS Rm 216** | Office hours: Thursdays 5-6pm | Course Website: https://learn.bu.edu/
**Faculty: Hong Pan, PhD** | Lecturer, Computer Science Department
Email: hongpan@bu.edu | Mobile: (917)439-2996 | URL: www.bu.edu/csmet/profile/hong-pan
**TA: Sriya Adiraju** | Email: sriya17@bu.edu | **TA: Zeyu Cai** | Email: zeyucai@bu.edu
Recitation hours**: TBD**

## Table of Contents

## Course Description

This course covers foundations of machine learning, regression, and classification. 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 these methods and interpret their outputs and how to visualize the 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 to help students to learn when and how to deploy different methods.

**Prerequisites:** CS546 (Introduction to Probability and Statistics) and CS544 (Foundations of Analytics and Data Visualization) or equivalent background.

# Learning Objectives

By successfully completing this course you will be able to:

- You will learn basic blocks of regression and classification
- Understand the basics of machine learning
- Summarize and present data in meaningful ways
- Select the appropriate analysis depending on research questions at hand
- Form testable hypotheses that can be evaluated using statistical analyses
- Understand and verify the underlying assumptions of a particular analysis
- Effectively and clearly communicate results from analyses performed with others
- Conduct, present, and interpret data analyses using R
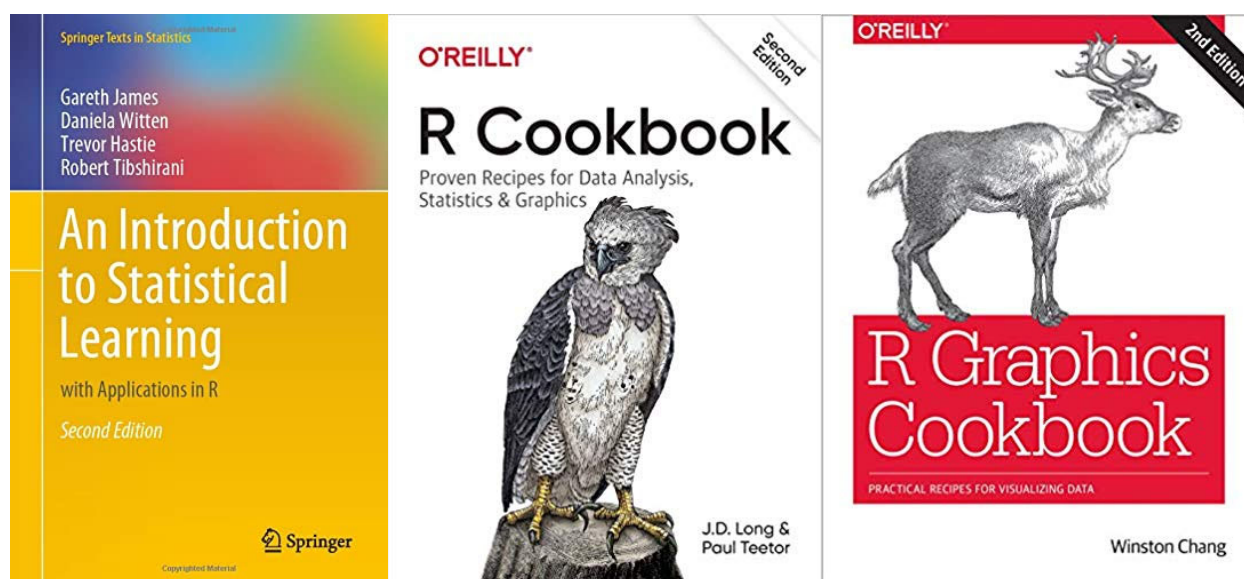
# Textbook

The following book is the textbook for the course. It is highly recommended to use the textbook alongside of the course material.

- **An Introduction to Statistical Learning with Applications in R, second edition (2021)** ; by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Published by Springer. The book has been made available online for free at https://statlearning.com. You can also purchase a hard copy.

There will be no specific reading assignments from the Textbook.

# Reference Books

The following books are not required for the course. They can be used as a reference and help with the assignments and term project.

The following books are excellent supplemental texts for R that you may want to review as we go through the course.

- 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 Reference Books**

- 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.
- "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)

**Boston University** Metropolitan College

# Class Policies

- **Assignment Completion & Late Work** – all the assignment must be submitted in person or electronically on Blackboard. **No late work** will be acceptable.
- **Laptop Requirement –** Students should have a personal laptop. We will use laptops in classroom to write R programs. You will need a laptop in quizzes and the final exam as well. Please have your laptop **FULLY CHARGED** before coming to the classroom every Thursday!
- This is an on-campus class. Class attendance is **mandatory** and will be recorded. Unexcused absences could result in lower grades.
- **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
- Class participation (5%)
- Quizzes (20%)
- Assignments (25%)
- Final project (20%)
- Final exam (30%)

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

**Homework Assignments**

There will be **6 Module Homework Assignments**, assigned 1 week before the due date and due 1 hour before 6 corresponding Module Quizzes, focused on applying theory learned in the class to analyze a data set in R. Assignment submissions should be in a single **PDF** file. The R code used to generate your results should be appended to the end of your assignment.

**Quizzes**

- There will be **6 Module 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 materials.
- There will be **Weekly Short Quizzes** at the beginning of the class time (except for the 6 weeks of Module Quizzes), and **In-Classroom Practices** which will be submitted at the end of the class time.

**Project**

The project is open ended and the topics can be chosen by students. In this project, students will frame and solve problems using quantitative capabilities of Statistical Learning with R. Students will present their projects in the final week of the course.

**Final Examination**

The final exam will be comprehensive and will cover material from the entire course.
The final exam will be <u>closed notes and closed book</u>.

# Course R Code Resources on Github

Our course R code examples are available on Github (https://github.com/kiat/R-Examples ).
- You don't need to register the Github account to access the course R code resources.
- You can download them all as a single zip file.

# Running R Programs

We will be using RStudio IDE (Integrated Development Environment) and R. We have these installed in our virtual lab. **MET Virtual Labs (VLAB)** provide students with all required software. Most of the examples presented in class will be run in this environment. You can familiarize yourself with the virtual labs with the information from our website:
- For the latest instructions to access:
https://www.bu.edu/metit/services/client-technology/virtual-lab/
- For saving files from the virtual labs:
https://www.bu.edu/metit/services/client-technology/virtual-lab/saving-documents/

**Boston University** Metropolitan College

# Study Guide

**Module 1: Describing and Interpreting Data**
  **Required Reading:** Module 1 Lecture Notes
  **Optional Reading:**
  - Teetor, Sections 2.6, 2.13, 3.1, 3.6, 3.9, 8.9, 8.10, 8.11, 9.1, 9.2, 9.5, 10.9, 10.11, 10.16, and 10.18
  - Chang, Sections 1.3, 1.4, 2.3, 2.4, 2.5, 3.1, 6.1, 6.6, 13.16
  **Assignments:** Assignment 1
  **Assessments:** Quiz 1

**Module 2: Statistical Inference and Tests for Comparisons of Means**
  **Required Reading:** Module 2 Lecture Notes
  **Optional Reading:**
  - Teetor, Sections 8.9, 8.10, 9.8, 9.9, 9.15, 10.9, 10.10, and 10.17
  **Assignments:** Assignment 2
  **Assessments:** Quiz 2

**Module 3: Correlation and Simple Linear Regression**
  **Required Reading:** Module 3 Lecture Notes
  **Optional Reading:**
  - Teetor, Sections 2.6, 9.17, 10.1, 10.6, 11.1, 11.3, 11.4, and 11.13
  **Assignments:** Assignment 3
  **Assessments:** Quiz 3

**Module 4: Regression Diagnostics and Multiple Linear Regression**
  **Required Reading:** Module 4 Lecture Notes
  **Optional Reading:**
  - Teetor, Sections 11.2, 11.8, 11.10, 11.11, 11.14
  - Chang, Section 5.13
  **Assignments:** Assignment 4
  **Assessments:** Quiz 4

**Module 5: Analysis of Variance**
  **Required Reading:** Module 5 Lecture Notes
  **Optional Reading:**
  - Teetor, Sections 11.20, 11.21, 11.22

**Assignments:**   Assignment 5
**Assessments:**   Quiz 5

**Module 6: Tests for Comparisons of Proportions and Logistic Regression**
   **Required Reading:**   Module 6 Lecture Notes
   **Optional Reading:**
   • Teetor, Sections 9.11, 9.12, 9.18, and 13.7
   **Assignments:**   Assignment 6
   **Assessments:**   Quiz 6

**Final Project:**
   **Initial Proposal: 1**-2 double-spaced pages, due Thursday, March 16 5pm ET
   **Final Term Paper:**  At least 8 double-spaced pages, due Thursday, April 27 5pm ET
   **Final Term Project Presentation:** 2-minute elevator pitch style presentation, due Thursday, May 4 6pm ET

# Course Survival Guide

- Attend every class in person and arrive on time; Be attentive and ask questions; Keep your phone off and away; Study the course materials posted on Blackboard BEFORE coming to class and review them AFTER; Manage and plan for your Homework Assignment deadlines, for no late work will be accepted.
- Be proactive on Blackboard **Class Discussion** Board, asking questions on the Board during and after the class time (some of them will be answered during the class time, and others will be answered outside of class or during the weekly recitation hour), utilizing Blackboard **Water Cooler** Board to form study groups to discuss concepts.
- Attend your weekly recitation hour, use the opportunity to ask questions, go over complicated topics and review for quizzes.
- Consider learning statistics and R programming as like learning a foreign language, which requires: Time commitment (**at least 6 hours per week** outside of class) and practice regularly (**15 minutes per day** will make a big difference within the short period of a semester).
- Keep yourself as healthy as possible to foster your best learning.

# Calendar

| Month | Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|---|---|---|---|---|---|---|---|
| **Jan 2023** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| | 29 | 30 | 31 | 1 | 2 | 3 | 4 |
| **Feb 2023** | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| | 26 | 27 | 28 | 1 | 2 | 3 | 4 |
| **Mar 2023** | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| | 26 | 27 | 28 | 29 | 30 | 31 | 1 |
| **Apr 2023** | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| | 30 | 1 | 2 | 3 | 4 | 5 | 6 |
| **May 2023** | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| | 28 | 29 | 30 | 31 | 1 | 2 | 3 |
| **Jun 2023** | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| | 18 | 19 | 20 | 21 | 22 | 23 | 24 |

**Legend:**

**Class Day | Weekly Short Quiz Due at 6pm**

**Homework Assignment Due at 5pm**

**Module Quiz Due at 6pm**

**Weekly Recitation Zoom Hour**

**Term Project Presentation and Final Review**

**Final Exam**

**Final Project:**

**Initial Proposal**: 1-2 double-spaced pages, due Thursday, March 16 5pm ET

**Final Term Paper**: At least 8 double-spaced pages, due Thursday, April 27 5pm ET

**Final Term Project Presentation**: 2-minute elevator pitch style presentation, due Thursday, May 4 6pm ET

**Spring Break, Yah!**

# Academic Conduct Policy

Please visit Metropolitan College's website for the full text of the department's *Academic Conduct Code* (https://www.bu.edu/met/current-students/academic-policies-procedures/ ).

**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.**