**Data Analysis and Visualization**

MET-CS 555

On Campus (PSY B51)

**Instructor Name:** Vineet Raghu

**Email:** vraghu@bu.edu

**Office hours:** by appointment

**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 applications
* Summarize and present data in meaningful ways
* Select appropriate statistical tools for 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

**Books**

The following textbook is required, and is freely available online at the following link: <https://rc2e.com/>

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

The following books are excellent reference material for use during the course and for the future

* Andy Field, Jeremy Miles, and Zoe Field. (2012). Discovering Statistics using R. Publisher: SAGE Publications Ltd. ISBN-13: 978-1446200469
* Chang, W. (2013). R graphics cookbook. Sebastopol, CA: O’Reilly. ISBN 9781449316952.
* Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (2013). An introduction to Statistical Learning with Applications in R. Springer.
* <https://www.openintro.org/stat/> Free PDF download with R tutorials

**Courseware**

The website for this course is available on Courseweb. Please contact the instructor if you are unable to access the course.

The website will be updated with example R notebooks and presentations from lecture.

**Class Policies**

1. **Attendance & Absences** – Attendance and participation in class is expected. Advanced notice should be sent to the instructor to miss a class.
2. **Assignment Completion & Late Work** – All assignments should be submitted on the course webpage on time.

Late work without a valid excuse will result in a penalty. Submissions **24 hours late** will receive a **10% penalty**. Submissions **48 hours late** will receive a **20% penalty.** Assignments turned in after this time will not be accepted.

There are no makeup dates for midterm and final exams. If you have an emergency commitment on these dates, an alternate date can be scheduled with written permission (via email) from the instructor.
3. **Laptop Requirement –** Students must have a personal laptop. Laptops will be used in class to write programs in R and are required for the exams.
4. **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**

|  |  |
| --- | --- |
| **Deliverable** | **Weight** |
| Homework Assignments | 30% |
| Quizzes | 10% |
| Midterm Exam | 20% |
| Term Project | 10% |
| Final Exam | 30% |

**Homework Assignments**

There will be six homework assignments that require you to analyze data using R and to present that data in a written report. R code used to generate the report should be included in the assignment. Assignments will be due the **Friday prior** to lecture. Assignments should be submitted as a single **PDF**.

**Quizzes**

Six quizzes will evaluate understanding of concepts presented in the previous weeks’ lectures. Quizzes will be held on the course website and will be due by the start of class.

**Midterm Exam**

The midterm exam will be an online evaluation of the first half of the course material. More details will be given later in the course.

**Class Meetings, Lectures & Assignments**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Topic** | **Quiz** | **Assignments Due** |
| January 27 | * Fundamentals of statistics
* Summarizing quantitative and qualitative data
 |  |  |
| February 3 | * Normal Distribution
* Sampling
* Central Limit Theorem
 |  |  |
| February 10 | * Statistical Inference
* Defining and evaluating statistical hypotheses
* Test statistics and p-values
 | Quiz 1 Due | Assignment 1 Due (February 7 11:59 PM) |
| February 18  | Class held on Tuesday due to President’s Day Holiday* Significance tests and confidence intervals
* Inference about a population mean
* Two-Sample Problems
 |  |  |
| February 24 | * Scatterplots and correlation
 | Quiz 2 Due | Assignment 2 Due (February 21 11:59 PM) |
| March 2 | * Simple Linear Regression
* F-test for Simple Linear Regression
* t-test for Simple Linear Regression
 |  |  |
| March 9 | NO CLASS (Spring Recess) |  |  |
| March 16 | * Residual plots
* Outliers and influence points
* Assumptions of Least-Squares
 | Quiz 3 Due | Assignment 3 Due (March 13 11:59 PM) |
| March 23 | * Multiple linear regression
* F-test and t-test for multiple regression
* Cautions about regression
 |  | Midterm Exam Due (March 20 11:59 PM) |
| March 30 | * Analysis of Variance (ANOVA)
* F-test for ANOVA
* Evaluating group differences
 | Quiz 4 Due | Assignment 4 Due(March 27 11:59 PM0 |
| April 6 | * Issues with multiple comparisons
* Assumptions of ANOVA
* One-Way analysis of covariance
* Two-Way analysis of variance
* Two-Way analysis of covariance
 |  |  |
| April 13 | * Significance tests for proportions
* Confidence intervals for a proportion
 | Quiz 5 Due | Assignment 5 Due(April 10 11:59 PM) |
| April 22 | Class held on Wednesday due to Patriots’ Day Holiday* Two-Sample tests for proportions
* Confidence intervals and significance test for differences in proportions
* Effect Measures
* Logistic Regression
* Area under the ROC curve (AUC)
 |  |  |
| April 27 | Final Review Session | Quiz 6 Due | Assignment 6 Due(April 24 11:59 PM) |
| May 5-9 | Final Exam (Date TBA) |  |  |