Carnegie Mellon University
Heinz College

95-868: Exploring and Visualizing Data, Summer 2021 Mini 5

Lecture Times

Lectures: 1 hour pre-recorded lectures twice per week

Instructors and Office Hours

Professor:
Donald P. Taylor, PhD, MBA, CLP
dontaylor@cmu.edu

Teaching Assistants:

Mr. Saharsh Agarwal
saharshagarwal@cmu.edu

Ms. Katherine Diaz
kdiaz@andrew.cmu.edu

Out-of-classroom Assistance: Email the instructor of TAs at least 48 hours in advance to schedule a Zoom session. Please also use the Discussion Forum on Canvas for asking general questions, questions related to each week's topic and homework, and the final project. The use of the Canvas Discussion Forum is important because that way all students benefit from the dialogue.

Course Description

This distance-learning course covers the fundamentals of statistical exploration and visualization of data. We will fit models and produce specialized graphs to explore data in a detailed and statistics-oriented manner. This course also serves as an introduction to R, a widely used statistical programming language.

In this class, students will learn:

1. How to use R to perform basic data manipulation such as filtering, aggregating, and organizing data sets, and for production of graphics

2. How transformations, model fits, and residuals can be used to explore and check statistical assumptions about data

3. How simulation can be used to explore questions of model fit and statistical significance

Course policies

Prerequisites

A first course in statistics is required, such as either 95-796 or 90-711.

Computer policy

You will need a computer capable of installing R and RStudio

Textbooks

Helpful references for using R. These aren't necessary but might be helpful references for you in the future.

1. R for Data Science, by Hadley Wickham (try this one first, it's the most modern)

2. R Graphics Cookbook, by Winston Chang

3. R for Everyone, by Jared Lander

4. R Cookbook, by Paul Teetor

Finally, if you are interested in a deeper discussion of some of the statistical methods presented in class, you may want to check out:

- Visualizing Data, by William Cleveland (caution: it is dry)

Coursework and Grading

Your grade in this course will be based on 5 homework assignments and a mini project.

- Homework (60%)
  - There are 5 homeworks, equally weighted towards your grade.
  - Homework should be submitted online via Canvas, by 11:59 PM EST on the scheduled due date.
  - Each student has 5 late days in total – these late days can only be used across the 5 homeworks. You may use them at your discretion, to cover travel for interviews, illness, or general business. Otherwise, late homework will not be accepted. **Note – late days cannot be applied to the Mini-Project.**
- Mini-Project (40%)
  - We will give you a data set and one or more questions to explore.

- The project will cover one or more aspects of the course (so not everything you have learned will necessarily be relevant).
- You will be responsible for structuring the analysis yourself and deciding what tools should be used.
- You will also need to present your analysis and results in a clear and concise manner in the form of a final report.
- **No late days may be applied to the Mini-Project.**

Grading

We follow the CMU Heinz Grading scale that can be found at the below link.

https://www.heinz.cmu.edu/heinz-shared/_files/img/student-handbooks/2017-2018-heinz-college-handbook.pdf

A+	Exceptional	98% - 100%
A	Excellent	94% - 97%
A-	Very Good	90% - 93%
B+	Good	86% - 89%
B	Acceptable	82% - 85%
B-	Fair	78% - 81%
C+	Poor	74% - 77%
C	Very Poor	70% - 73%
C-	Minimal Passing	66% - 69%
D,R	Failing	<66%
I	Incomplete	

Homework Grading Rubric (20 points total)

- 1 pts: if the code compiles
  - 1/1 if the grader runs "knit HTML" and an HTML document is returned instead of errors
  - 0/1 if we ask you to resubmit because the homework did not compile the first time
  - if you do not resubmit in 36 hours: homework is LATE.

Assuming the HW eventually is compile-able:

- 17 pts: correct solution

        - Full credit for each problem that is 100% correct or has inconsequential errors
        - Partial credit for each problem with errors
        - No credit for each problem that is not attempted
- 2 pt: readability
    - 2/2 if the homework is well written:
        - variables are well named
        - avoids needless replication
        - good commenting within R code
        - reasonable coding style
        - non-code writing is clear and to the point
    - 1/2 if exposition is pretty good but doesn't deserve full credit
    - 0/2 if there are significant coding or writing issues, and we don't understand what you are doing

Professor & Teaching Assistant Communication

Questions should be sent at least 48 hours before any homework or final project deadline to assure a timely response. Please include the course code 95-868 in the subject line of your emails. There is also a Canvas forum created for questions and answers that will be actively moderated to ensure only clarifications to the assignments are discussed and not answers to questions. **We encourage you to use the Canvas Discussion forum as the primary medium to ask questions**. Please be conscientious about not posting any answers to assignments on the forum (or anywhere) in order to maintain academic integrity.

Collaboration

You are encouraged to discuss general approaches and clarification questions with your fellow students. However, you should do your homework yourself.

- Do not look at (or copy) another student's homework.
- Do not copy from another student's homework.

If you receive any help from another student or outside the class (such as stackexchange or other forums or websites), you must clearly identify where you received help. The expectation is that your grade must reflect the work that you alone did.

Tentative Schedule

Week 1:
Introduction to R, Rstudio, and RMarkdown
Data cleaning and aggregation

Week 2:

Graphics parts 1 and 2

Week 3:
Averages and sample sizes, part 1 (p-values)
Averages and sample sizes, part 2 (confidence intervals)

Week 4:
Mon: Univariate distributions part 1 (quantiles and QQ plots)
Wed: Univariate distributions, part 2 (residuals and transforms)

Week 5:
Functions of one variable, part 1 (splines and cross-validation)
Functions of one variable, part 2 (quantile regression)

Week 6:
Interactions and multivariate models, part 1 (interactions)
Interactions and multivariate models, part 2 (stepwise variable selection)
Logistic regression and generalized linear models
Correlation matrices, clustering, and heatmaps