Course Title:	Programming R for Analytics
Course Number:	94-842
Section:	A1 (Fall 2017 Mini 1)
Meeting times:	M/W 9 – 10:20 am
Meeting location:	HBH 1204
Instructor:	Mike Blackhurst, mblackhurst@gmail.com
Teaching Assistants:	Andres Salcedo
	Ankit Gupta
Office hours:	Locations and times TBD

MOTIVATION, SCOPE, AND PURPOSE OF THE COURSE MATERIAL

The volume, variety, and velocity of machine-readable data are increasing and, in turn, data are progressively informing decisions made by individuals, businesses, non-profits, and governments. *R* is a free computing environment and programming language designed specifically for data science that has significantly grown in user base, capability, and end use applications. The purpose of this course is to develop basic, foundational data science methods and R programming skills that can prepare students for a variety of career paths involving analytics.

LEARNING OBJECTIVES

The learning objectives of this course are as follows: (1) read and write information into and out of R; (2) develop, apply, and document R code in a preferred style that applies operators and functions to data; (3) write and use your own R functions; (4) design and develop visualizations for rectangular data containing different levels of measurement; (5) develop summary statistics of data in R; (6) formulate, perform, and interpret simple hypotheses tests in R; (7) perform and interpret simple OLS linear regression models in R; (8) execute sampling in R; (9) communicate analytical methods executed in and results derived from R code.

PRIOR KNOWLEDGE

No prior programming knowledge is assumed. Students should have some understanding of probability distributions and previous experience with rectangular data (such as data typically stored in spreadsheets).

SUMMARY OF COURSE ACTIVITIES

Course activities will be lectures, which will primarily be demonstrations connecting the learning objectives to the respective R programming skills, and in-class labs, which will typically be done in groups. Students will often be asked to share their approach to completing the labs with the class. The course ***may*** include a recitation and online videos demonstrating R programming techniques, depending on developing needs throughout course.

COMMUNICATION

The course will use Piazza for communicating course-wide information, including posing and answering questions that may be relevant to multiple students. For other communications, such as requesting an excused absence, students should feel free to directly email me.

The course will use Canvas to record student grades and disseminate assignments, data, in-class labs, lecture materials, and other similar content. Students will submit assignments on Canvas.

ATTENDANCE, PARTICIPATION, & COMPUTERS & CELLPHONES POLICY

Class participation points are given to recognize your engagement with both lectures and in-class labs. For each class, students are required to sign an attendance sheet indicating their arrival time. All students involved in fabricating signatures will receive a zero grade for course participation. Students will be allowed to arrive 5 minutes late to class without penalty once. Each additional arrival beyond 5 minutes will reduce class participation scores by 1 point. Excused absences will be granted by permission of the instructor. I may petition appropriate Carnegie Mellon staff and/or administrators to drop students from the course who have excessive absences.

Students will need to use their computers during the in-class labs. Outside of labs, I will not monitor the use of laptops and cell phones during class. However, it should be emphasized that research has shown that performance in negatively affected when students use computers in class for non-classroom activities. *More importantly, students will lose 1 class participation point each time their computer or cell phone distracts others in the classroom.*

ACADEMIC INTEGRITY AND COLLABORATION

Students are encouraged to ***discuss*** assignments with your peers and ***review*** resources that assist with assignment completion. However, students must submit their own code and acknowledge any resources, including peer assistance, as comments in code. In the context of coding, students are to individually write and submit their own code. You are allowed to copy and modify code presented in lectures and as part of labs, as these are provided as instructional resources. This is materially different than copying your peers' code.

The following passage from the University of Washington[1] provides guidelines for distinguishing between collaboration and plagiarism.

> *"[It is] important to make sure that the assistance you receive consists of general advice that does not cross the boundary into using code or answers written by someone else. It is fine to discuss ideas and strategies, but you should be careful to write your programs on your own."*
>
> *"You must not share actual program code with other students. In particular, you should not ask anyone to give you a copy of their code or, conversely, give your code to another student who asks you for it; nor should you post your solutions on the web, in public repositories, or any other publicly accessible place. [You may not work out a full communal solution on a whiteboard/blackboard/paper and then transcribe the communal code for your submission.] Similarly, you should not discuss your algorithmic strategies to such an extent that you and your collaborators end up turning in [essentially] the same code. Discuss ideas together, but do the coding on your own."*
>
> *"Modifying code or other artifacts does not make it your own. In many cases, students take deliberate measures -- rewriting comments, changing variable names, and so forth -- to disguise the fact that their work is copied from someone else. It is still not your work. Despite such cosmetic changes, similarities between student solutions are easy to detect. Programming style is highly idiosyncratic, and the chance that two submissions would be the same except for*

[1] https://www.cs.washington.edu/students/policies/misconduct

changes of the sort made easy by a text editor is vanishingly small. In addition to solutions from previous years or from other students, you may come across helpful code on the Internet or from other sources outside the class. Modifying it does not make it yours."

"[I] allow exceptions in certain obvious instances. For example, you might be assigned to work with a project team. In that case, developing a solution as a team is expected. The instructor might also give you starter code, or permit use of local libraries. Anything which the instructor explicitly gives you doesn't normally need to be cited. Likewise, help you receive from course staff doesn't need to be cited."

Students are not to collaborate on the final project.

Assignments that do not meet these criteria will receive a zero score and be reported to appropriate Heinz/CMU staff. Egregious instances of plagiarism may result in a failing grade for the course.

STATEMENT ON STUDENT WELLNESS[2]

I care about your wellness. Do your best to maintain a healthy lifestyle this semester by taking care of your physical, mental, and emotional help, including making time to relax. This will help you achieve your goals, cope with stress, and practice the balancing you will need to maintain sustained success at Carnegie Mellon and beyond.

All of us benefit from support during times of struggle. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is almost always helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website athttp://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

ACCOMMODATIONS[3]

If you have a disability and have an accommodations letter from the Disability Resources office, discuss your accommodations and needs with me as soon as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

ASSESSMENTS, ASSESSMENT SCHEDULES, AND ASSESSMENT FORMATS

Participation (10% of final grade). Attendance and participation in class labs constitute your grade for participation. There are no separate deliverables for the labs graded independent of classroom learning. Your grade for participating in labs is made at the discretion of the instructor. Examples of participation worthy of a perfect score include consistently and actively working on each lab when assigned, asking questions when you get stuck, contributing to the solution when

[2] Statement modified from https://www.cmu.edu/teaching/designteach/syllabus/checklist/studentwellness.html

[3] Statement modified from http://www.cmu.edu/teaching/designteach/syllabus/newcourse/coursepoliciesandstatements.html

working in groups, and sharing your completed work with others when asked. Points can be lost if you are distracted by your phone, email, or other non-course content on your computer, unwilling to participate when working in groups (including insisting on doing things "your way"), or unwilling to ask questions as necessary. The labs are designed to be fully completed in class, but completion is not necessary for full credit. You will receive full credit for participation early in the semester and docked points as needed.

5 x quizzes (10% of final grade). To reward students that keep up with class materials, students will be given 5 quizzes on Wednesday during weeks 1-5. Students that keep up with the readings and attend lectures should be able to get full credit for quizzes. Each quiz will be worth 2 points. Quizzes will be given using online survey software (e.g., Google Forms).

5 x assignments (50% of final grade). Students are expected to complete 5 assignments at 10 points each. Assignments are to be provided in RMarkdown (.Rmd) format. I have provided a RMarkdown template for assignment 1 ("Template_Assingment1.Rmd") on Canvas that you can use as a reference. File names for assignments must follow the convention "Lastname_Firstname_Assignment#.Rmd." For example, John Smith will call his completed assignment 1 "Smith_John_Assignment1.Rmd." The grading rubric for assignments will be 1 point for proper knitting, 1 point for style, and 8 points for correctness, made at the discretion of the teaching assistants. If your code does not knit, you will be given 24 hours to correct your code from the time the TA notifies you. ***IMPORTANT:*** *Some assignments will require you to import data stored locally on your machine. For it to knit properly for others, you will need to comment out the code using for importing and TA's will need to edit this code to reference their local file folder pathname used for assignment data.* Unless otherwise noted, assignments are handed out on Wednesdays and due by midnight the following Wednesday. The TA's and instructor will make all reasonable efforts to return graded assignments within one week of their due date.

Final Project (30% of final grade). Students will be asked to investigate a current policy issue using a raw, publically available data. The final report will be graded using 5 criteria, each worth 6 points: (1) data exploration, transformations, and wrangling; (2) the design and presentation of appropriate figures and tables; (3) the selection and application of statistical analysis; (4) the quality of the narrative used to describe methods and findings; and (5) the style, clarity, and appropriateness of the R code used for the project. Students may choose any software to prepare their final project but must deliver their project in pdf format. The final project is due on Oct 20.

Assignments may include opportunities for ***extra credit*** at the discretion of the instructor.

LATENESS AND EXTENUATING CIRCUMSTANCES

Extenuating circumstances aside, late assignments and projects will not be graded. Extenuating circumstances, such as a death in the family or serious illness, will be evaluated on a case-by-case basis.

GRADING SCALE

A+	99.0-100%	B+	88.0-90.9%	C+	78.0-80.9%
A	94.0-98.9%	B	84.0-87.9%	C	74.0-77.9%
A-	91.0-93.9%	B-	81.0-83.9%	C-	71.0-73.9%

RESOURCES

REQUIRED (all free or selections posted on Canvas)

Garrett Grolemund and Hadley Wickham, R for Data Science. (Available for free at http://r4ds.had.co.nz/data-visualisation.html)

Morgan MG, Henrion M, Small MJ. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, 1990. (on Canvas)

Harrison, Robert L. "Introduction To Monte Carlo Simulation." AIP Conference Proceedings 1204 (January 5, 2010): 17–21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2924739/. (Subject to change if better materials are found.)

Pennsylvania State University. 2017a. Stat 461: Analysis of Variance. https://onlinecourses.science.psu.edu/stat461/ (subject to change)

Pennsylvania State University. 2017b. Stat 501: Regression Methods. https://onlinecourses.science.psu.edu/stat501/ (subject to change)

R Styles

adv-r.had.co.nz/Style.html

google.github.io/styleguide/Rguide.xml

SUPPLEMENTAL

R Operators

https://stat.ethz.ch/R-manual/R-devel/library/base/html/Arithmetic.html

http://www.statmethods.net/management/operators.html

dplyr tutorial (http://genomicsclass.github.io/book/pages/dplyr_tutorial.html)

R cheat sheets (https://www.rstudio.com/resources/cheatsheets/)

Shortcut keys (https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts)

Many of the example data sets we will use come from Wooldridge, Jeffrey M. *Introductory Econometrics: A Modern Approach*. 6 edition. Boston, MA: South-Western College Pub, 2015.

EXPECTED SCHEDULE (SUBJECT TO CHANGE)

Class (Date)	Topics	Readings	Assignment=A Quiz=Q
1, Aug 28	Introduce R/ R Studio Levels of measurement Code styles R Markdown		A1 (out)
2, Aug 30	Visualizing data	G&W 2-4 See "R styles" under "Resources"	Q1
3, Sep 6	Data transformation and exploration	G&W 5-8	A1 (due), A2 (out)
4, Sep 11	Data cleaning part 1	G&W 10-12 (skim 11)	Q2
5, Sep 13	Data cleaning part 2	G&W 13-14	A2 (due) A3 (out)
6, Sep 18	Strings, factors, dates/times	G&W 14-16 (select parts TBD)	Q3
7, Sep 20	Pipes and functions	G&W 17-18	A3 (due) A4 (out)
8, Sep 25	Iteration, data communication, and formatting tables	G&W 19	Q4
9, Sep 27	Distributions, sampling, uncertainty, and variability analysis	M&H 4	A4 (due)
10, Oct 2	Hypothesis testing	PSU 2017a Lessons 1-6	A5 (out) Q5
11, Oct 4	Regression part 1	PSU 2017b Lessons 1-4	
12, Oct 9	Regression part 2	PSU 2017b Lessons 5-9	
13, Oct 11	TBD	None	A5 (due) TBD
14, Oct 16	Overview	None	
Oct 20			Projects due by midnight

REFERENCES

Some of the syllabus and course materials are modified from Chouldechova 2017.
http://www.andrew.cmu.edu/user/achoulde/94842/index.html