LIS 4370
Introduction to R for Data Science Professionals
3 Credits
College of Arts and Sciences
School of Information

Semester: Fall 2020
Online Class
Instructor: Dr. Alon Friedman
Office Location: School of Information Room # 2026
Office Hours: Tuesdays 9am to 4 pm via Microsoft Teams or Zoom
Phone Number: (813) 974-6840
Email: alonfriedman@usf.edu
GitHub: https://github.com/AlonFriedman01/R-class

I. Welcome!

R is a powerful programming language and software environment for statistical analysis and graphical data presentation. It is an open-source program, which means it is free to use, and its popularity reflects a shift in statistical computing and visualization output. The importance of R as a statistical language in today's data-rich environment is its utility in analyzing and visualizing large data sets and its robust community of users and developers. This course will teach you how to develop your own R package and to properly document your analysis and code.

II. University Course Description

This course aims to teach undergraduate students, across disciplines, how to use Open-Source R for data science professionals. It begins by providing a core foundation in the fundamentals of R programming workflow, data types, structures, and functions, and the syntax for coding in R. Students will become proficient in error handling/debugging to develop efficient and functional R programs. Because this is a undergraduate-level course, mastering the fundamentals just marks the beginning of the journey. Students will learn how to import a variety of data formats into R (using RStudio); to prepare or "tidy" datasets for analysis; to integrate R with relational databases (including MySQL queries); to select and apply an appropriate R package for a range of statistical methods to address specific types of data/research questions, and to create polished data visualizations using ggplot2 and other R packages to facilitate new data-driven insights. Students will even

learn to create and disseminate their own R packages, allowing them to develop specialized, advanced-level applications. The course will give special attention to writing and coding the R package documentation.

III. Course Prerequisites

Successful completion of any undergraduate course in statistics and/or programming.

IV. Course Purpose

This is not simply a "programming" course. Open-Source R is a specific software platform and language for data and information analysis. These skills are critical for information professionals and those in the social/behavioral and life/physical sciences as well. Over the semester, students will learn to write scripts for data analysis and compile a new R package (which is a collection of R functions, data, and compiled code in a well-defined format) to accomplish a specific data analytic objective. The course will also teach the student how to write supporting documentation for their code and package. This documentation is an important aspect of a good package. Without it, users won't know how to use your package.

V. Student Learning Outcome

This course provides students with the foundations necessary for understanding, applying, and extending open-source R for data analysis. Upon completing this course, students will be able to:

- Interpret R scripts and to utilize R packages to analyze data
- Define and critique R variables, functions, and sequences while writing supporting documents to support an R package
- Associate R technologies and its workflow with foundations of data science components (e.g., data abstractions, statistical/data mining, and visual encodings) by constructing an R package.

.

VI. Course Topics

- Introduction to R
- Basic foundations for R programming
- Understanding different data types in R
- Doing math in R
- R as an object-oriented programming
- String manipulation
- R package structure
- Visualization/graphics
- Debugging in R
- R markdown
- Documentation of R package
- Using Shiny
- Future of R

VII. Required Texts and/or Readings and Course Materials

- The Art of R Programming: A Tour of Statistical Software Design. Norman Matloff. No Starch Press San Francisco, CA (2011) ISBN 10: 1-59327-384-3
- R Packages, Organize, Test, Document and Share Your Code, Hadley Wickham. O'Reilly, Sebastapol, CA (2015) ISBN-13: 978-1491910597
- Advanced R, Second Edition. Hadley Wickham. CRC Press, Boca Raton, FL (2019). ISBN: 978-0815384571

VIII. Supplementary (Optional) Texts and Materials

- Provided in the course itself

IX. Grading Scale

Grading Scale (%)	
94-100	A
90 – 93	A-
87 – 89	B+
84 – 86	B
80 – 83	B-
77 – 79	C+
74 – 76	C
70 – 73	C-
67 – 69	D+
64 – 66	D
60 – 63	D-
0 – 59	F

X. Grade Categories and Weights

Your grade will be based on the following:

Assessment	Percent of Final Grade
Blog	5%
Weekly Homework Assignment	30%
Student Peer Reviews	15%
Major/Final Project	50%

XI. Project Assignments

1. **Blog (5% of the grade)**
   Each student in the class will need to create his/her own blog. There are several different sites offering free blog services, but one of the easiest to use is WordPress. You can also create your blog on the USF system. Your blog postings need to reflect your diary of your own experiences with advanced statistics. You are expected to complete at least one blog posting a week.
2. **Weekly Homework Assignment (30% of the grade)**
   Each week you will be assigned homework that will consist of 12 exercises that will summarize the module we covered. Completing or attempting the homework is very important to succeed in this class because it gives you an opportunity for practice and application. It is expected that mistakes will occur in practice; therefore, incorrect (not incomplete or partially complete) answers on homework problems will not result in a significant penalty. The lowest homework grade will be dropped from the calculation of the final grade.
3. **Student Peer Reviews for proposal and final projects (15% of the grade).**
   Student code review aims: improving the code, finding bugs, anticipating possible bugs, checking the clarity of the code, and checking for consistency with the project's style standards. In this class, you will be given two opportunities to review your peers' work. The rubric score we will use in this class consist of numeric evaluation and open remarks. Your participation in this process will help you to get feedback beyond your professor's point of view on your work.
4. **Major/Final Project (50%)**
   The aim of this project is to experience data analysis where you will use the statistical methods taught in this course to come up with some objective findings. You will collect the data from your own choice of data repositories and conduct inferential analysis as discussed in the class. The Final Project will be graded by the following criteria:
   - o Appropriateness, thoroughness, and accuracy of code organization 40%.
   - o Effectiveness of writing communication for your work 40%. Including documentation workflow, terminology and index of your functions using vignettes format.

XII. Grade Dissemination

Graded assignments and projects in this course will be returned individually. You can access your scores at any time using "Grades" in Canvas.

XIII. Course Schedule

Date	Topics	Topics to be Discussed in Class	Reading & Homework Assignments
Module # 1	First day of class Introduction to R	• Installing R and RStudio on your machine • Retrieving R packages • The R console • Review of RStudio interface • Numbers, characters, and composites	Required Readings: • Matloff Chapters 1-2 (what is vector etc) • Wickham, H. (2015) R Packages Chapter 1 • Paradis, E. (2000). R for Beginners. • RStudio Cheat sheets – See https://rstudio.com/resources/cheat sheets/ • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Open accounts on GitHub, Wordpress and download R and Rstudio. • Share those links on Canvas.
Module # 2	R Foundations and Style	• Basic R data types and operations • Numbers, characters, and composites • Vectors, creating sequences, common functions • Importing data into R • Simple summaries or categorical and continuous data • R style basics • * Introduction to GitHub - Use RStudio to connect to your GitHub repository • View your data and R code repository via your browser.	Required Readings: • Matloff Chapters 3-4-5-6 • Wickham, H. (2015) R Packages Chapters 3 & 13 • Joan Clavdrol Writing functions in R- Data Camp Notes https://rpubs.com/JoanClaverol/493717 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Submit all 5 code exercises assigned for Module 2 on your GitHub account. Each exercise as separate file.

			• Post your written summary result of those exercises on your blog with a link to your GitHub.
Module # 3	Data Structures in R	• Vector Data • Attributes • Data Array • Data Matrix - Data.frame • Analyzing data using data. frame • Introduction to "tidy" package • Connecting via RStudio to different databases (MySql, SQL etc)	Required Readings: • Wickham, H. (2019). *Advanced r*. Chapters 1-4 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 5 code exercises assigned for Module 3 on your GitHub account. Each exercise as separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub
Module # 4	R Programming Structures	• Loops • Arithmetic argument • Boolean operators • Return Values • Functions as objects • Input/Output functions • Replacement functions	Required Readings: • Matloff Chapter 7 • Wickham, H. (2019). *Advanced R*. Chapter 7-9 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. <u>Homework Exercises</u>: • Post all 4 code exercises assigned for Module 4 on your GitHub account. Each exercise as separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub

Module # 5	R: The Mathematical & Statistical Machine	• Probability analysis
		• General statistics analysis
		  - Descriptive, Sample
		  - Confidence interval estimation
		  - Hypothesis testing
		  - Two sample tests and Chi Square Test
		• Linear regression and ANOVA analysis
		• Time Series analysis
		• Multivariate analysis
			Required Readings:
			• Matloff Chapters 8
			• Supplemental Recommended Readings on the module's topic will be posted on Canvas.
			
			Homework Exercises:
			• Post all 5 code exercises assigned for Module 5 on your GitHub account. Each exercise as separate file.

		• Data Modeling	• Post your written summary result of those exercises on your blog with a link to your GitHub
Module # 6	R Object-Oriented Programming	• Names, values, objects, and classes • Functionals operators • S3 vs S4 • R as a metaprogramming • Metaprogramming expression	Required Readings: • Matloff Chapter 9 • Wickham, H. (2019). *Advanced R*. Chapter 10 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 7 code exercises assigned for Module 6 on your GitHub account. Each exercise as separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub
Module # 7	String Manipulation	• String manipulation • Regular expression • Input and output • * Introduction to plyr package • Transform and summaries • Common functions: match_df, reply, split, alply, indexed_array	Required Readings: • Matloff Chapter 11 • Wickham, H. (2019). *Advanced R*. Chapter 10 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 2 code exercises assigned for Module 7 on your GitHub account. Each exercise as a separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub
Module # 8	The Structure of an R Package	• The philosophy of R package • Code organization - Source/files Packages - Binary packages • Code Distribution - Source/files packages - Binary packages - The package's library • Supporting Documentation - Metadata - Vignettes	Required Readings: • Wickham, H. (2015) R Packages Chapters, 2-3 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 2 code exercises code assigned for Module 8 on your GitHub account. Each exercise as a separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub

Module # 9	Visualization/Graphics	• ggplot2 • Animation • Colourpicker • ggvis • ggforce • Lattice • RGL • Lombardi's package	Required Readings: • Matloff Chapter 12 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 2 visualization code exercises code assigned for Module 9 on your GitHub account. Each exercise as a separate file. • Post your written summary result together with the visualization those exercises will generate on your blog with a link to your GitHub
Module # 10	Debugging and Code Review	• Debug function • Traceback function • Trace function • GitHub code review • Peer Review in R • Rubric function	Required Readings: • Matloff Chapter 13 • Wickham, H. (2015) R Packages Chapter 7 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 2 code exercises assigned for Module 10 on your GitHub account. Each exercise as a separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub
Module # 11	Guest speaker:		Homework Exercise: Submit your proposal for the final project via Canvas. This proposal must include a short description of your project, code together with R package standards (statistical formula, data and documentation).
Module # 12	R Markdown and Building an R Package	• Markdown syntax - Inline formatting - Block level elements - Math - HTML widgets - LaTeX • Developing your R package - System setup - Development workflows	Required Readings: • Wickham, H. (2015) R Packages Chapters 4-6 pp. 33-69 and pp. 81-117 • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises:

		- Licensing - Namespace - Compiled code • Writing text documents via vignettes	• Post all 2 code exercises assigned for Module 12 on your GitHub account. Each exercise as a separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub
Module # 13	Shiny	• Shiny package • Shine web server • Shiny IDE • File structure • Trigger and update • Create and delate values • Modularize reactions	Required Readings: • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 2 code exercises code assigned for Module on your GitHub account. Each exercise as a separate file. • Post your written summary results of those exercises on your blog with a link to your GitHub.
Module # 14	Summary Notes and the Future of R	• Review Unified Parallel and Distributed Processing • Review AI in R • Review Cloud computing in R • Quantum computing in R	Required Readings: • Supplemental Recommended Readings on the module's topic will be posted on Canvas. Homework Exercises: • Post all 2 code exercises assigned for Module 14 on your GitHub account. Each exercise as a separate file. • Post your written summary result of those exercises on your blog with a link to your GitHub
Final project	*Final Project is due May 7 **Peer review comments must be submitted by XXX		

* Note: The Schedule is subject to revision

XIV. Standard University Policies

Policies about disability access, religious observances, academic grievances, academic integrity and misconduct, academic continuity, food insecurity, and sexual harassment are governed by a central set of policies that apply to all classes at USF. These may be accessed at: https://www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx

XV. Course Policies: Grades

Late Work Policy:

In this class, there are no make-ups for in-class weekly assignments or final project. The final project turned in late will be assessed a penalty: a half-letter grade if it is one day late, or a full-letter grade for 2-7 days late. Final project will not be accepted if overdue by more than seven days.

Extra Credit Policy: This course will not offer extra credit.

Grades of "Incomplete":

The "I" grade may be awarded to a student only when a small portion of the student's work is incomplete and only when the student is otherwise earning a passing grade. The time limit for removing the "I" is to be set by the instructor of the course. For undergraduate students, this time limit may not exceed one academic semester. "I" grades not removed by the end of the time limit will be changed to "IF" or "IU," whichever is appropriate.

.

XVI. Course Policies: Technology and Media

Email:

You can email the instructor via USF email or via Canvas email system.

Canvas:

This course will be offered via USF's learning management system (LMS), Canvas. If you need help learning how to perform various tasks related to this course or other courses being offered in Canvas, please view the following videos or consult the Canvas help guides. You may also contact USF's IT department at (813) 974-1222 or help@usf.edu.

Course Blog:

This course required the student have a blog, to report the assignment and link to GitHub repository. Each student in the class will need to create his/her own blog. There are several different sites offering free blog services, but one of the easiest to use is WordPress. You can also create your blog on the USF system. Your blog postings need to reflect your diary of your own experiences with R programming. You are expected to complete at least one blog posting a week throughout the semester.

XVII. Course Policies: Student Expectations

Attendance Policy:

Students are expected to attend this online class' lectures, assignments, and guest speakers but attendance is only directly taken on the first day of class. The grading policy allows you to miss one assignment, due to illness or other emergencies, without penalty.

Course Hero Policy:

The USF Policy on Academic Integrity specifies that students may not use websites that enable cheating, such as by uploading or downloading material for this purpose. This does apply specifically to CourseHero.com – any use of this website (including uploading materials) constitutes a violation of the academic integrity policy.

Professionalism Policy:

University policies for classroom etiquette should be observed for any synchronous, interactive online sessions. Per university policy and classroom etiquette; mobile phones, iPods, etc. **must be silenced** during all classroom and lab lectures. Those not heeding this rule will be asked to leave the classroom/lab immediately so as to not disrupt the learning environment. Please arrive on time for all class meetings. Students who habitually disturb the class by talking, arriving late, etc., and have been warned may suffer a reduction in their final class grade.

End of Semester Student Evaluations:

Blog (5% of the grade)
Weekly Homework Assignment (30% of the grade)
Student Peer Reviews for selected modules (15% of the grade).
Major/Final Project (50%)
Total: 100% of the grade.

XVIII. Important Dates to Remember

Drop/Add Deadline:	Thu, XXX,
Spring Break:	Mon, XXX,
Submit the final project:	XXX,
Submit your peer comments	XXX,.