Term Information

Effective Term	Spring 2020

General Information

Course Bulletin Listing/Subject Area	Statistics
Fiscal Unit/Academic Org	Statistics - D0694
College/Academic Group	Arts and Sciences
Level/Career	Graduate, Undergraduate
Course Number/Catalog	5730
Course Title	Introduction to R for data science
Transcript Abbreviation	R for Data Science
Course Description	Introduces underlying concepts of the R programming language and R package ecosystem for manipulation, visualization, and modeling of data, and for communicating the results of and enabling replication of their analyses.
Semester Credit Hours/Units	Fixed: 2

Offering Information

Length Of Course	14 Week, 7 Week
Flexibly Scheduled Course	Never
Does any section of this course have a distance education component?	No
Grading Basis	Letter Grade
Repeatable	No
Course Components	Lecture
Grade Roster Component	Lecture
Credit Available by Exam	No
Admission Condition Course	No
Off Campus	Never
Campus of Offering	Columbus

Prerequisites and Exclusions

Prerequisites/Corequisites	1350, 1450, or 1550 or equivalent or permission of instructor
Exclusions	
Electronically Enforced	Yes

Cross-Listings

Cross-Listings

Subject/CIP Code

Subject/CIP Code	27.0501
Subsidy Level	Doctoral Course
Intended Rank	Junior, Senior, Masters, Doctoral

Requirement/Elective Designation

The course is an elective (for this or other units) or is a service course for other units

Course Details

Course goals or learning objectives/outcomes

- Understanding the basic concepts of the R programming language: expressions, objects, and data types
- Understanding a grammar of graphics and being able to construct visualizations of data
- Being able to efficiently manipulate and organize data
- Being able to apply different numerical techniques for exploring and summarizing data
- Understanding how data is represented, how to import data, the problems that may arise when importing data, and how to handle those problems
- Understanding different programming paradigms and abstractions in R (for example, piping, iteration, and functional programming), and being able to recognize where and when these concepts can be applied
- Being able to use different technologies for dynamically documenting, communicating the results of, and enabling replication of their data analyses with R and R Markdown

Content Topic List

- Data visualization
- Data types and representation
- Data frames and data manipulation
- Data summarization
- Data import and workflows
- Tidy data and relational data
- Functions and control flow
- Strings and factors
- Iteration and functional programming
- Dates and times
- Debugging and performance enhancement

Sought Concurrence

Yes

Attachments

- Re: Concurrence for STAT 5730 Introduction to R for Data Science.pdf: CSE Concurrence

  (Concurrence. Owner: Lee,Yoonkyung)
- syllabus.pdf: Syllabus

  (Syllabus. Owner: Lee,Yoonkyung)
- proposal.pdf: Rationale for the course proposal

  (Other Supporting Documentation. Owner: Lee,Yoonkyung)

Comments

Workflow Information

Status	User(s)	Date/Time	Step
Submitted	Lee,Yoonkyung	04/01/2019 05:47 PM	Submitted for Approval
Approved	Lee,Yoonkyung	04/01/2019 05:47 PM	Unit Approval
Approved	Haddad,Deborah Moore	04/01/2019 05:56 PM	College Approval
Pending Approval	Nolen,Dawn Vankeerbergen,Bernadette Chantal Oldroyd,Shelby Quinn Hanlin,Deborah Kay Jenkins,Mary Ellen Bigler	04/01/2019 05:56 PM	ASCCAO Approval

STAT 5730: Introduction to R for data science

Syllabus

Anonymous Instructor

Spring 2020

```
credit-hours: 2
format: lecture
preqrequisites: STAT 1350, 1450, 1550 or equivalent, or
                permission of instructor
```

1 Overview

R is a freely available statistical computing environment and programming language. It has become a dominant workhorse for modern statistical research and data analysis, and is being widely adopted in industrial data analytics as well. The primary goal of the course is to teach students how to use R for data analysis: both (1) efficient use of the R computing environment and (2) effective programming in the R language. This is not a course on general programming in the style of computer science. Our goal is to use R for data science. This involves manipulation, visualization, and modeling of data. So the class focuses on teaching the skills and underlying concepts of the R programming language and R package ecosystem that form a foundation for the computing required by those three tasks.

There are formal prerequisites for the course. This is a *statistics* course, so the examples and applications demonstrated in the class will be oriented towards data analysis and statistical endeavors. Basic numeracy and familiarity with statistics is expected for motivation and perspective. No programming experience is required.

2 Course materials & computing

- **Required reading**
  - (R4DS) Grolemund and Wickham (2016): *R for Data Science*. O'Reilly. ISBN: 9781491910382. (web: r4ds.had.co.nz). The web version of the book is can be accessed freely from any web browser. Electronic access to the print version of the book is available at https://www.safaribooksonline.com/library/view/-/9781491910382/?ar. **Note that the web and print versions have different chapter numbering.**

  - (HoPR) Grolemund (2014): *Hands-On Programming with R*. O'Reilly. ISBN: 9781449359089. Electronic access to the book is available at https://www.safaribooksonline.com/library/view/-/9781449359089/?ar
  - Print copies of both books can be purchased directly from oreilly.com. After visiting one of the above links, if you sign-up for a O'Reilly account with your OSU email address and install the appropriate app to your iOS or Android device (https://www.oreilly.com/online-learning/apps.html), you should be able to download the books for offline access.

- **Software**
  - R (www.r-project.org)
  - RStudio (www.rstudio.com)

You are expected to be able to access working installations of **current versions** of the required software. RStudio Server login access will be provided to students registered in the course at https://go.osu.edu/rstudio. This will allow you to access R via the RStudio IDE from any web browser. Alternatively, you can also install R and RStudio on your personal computer by downloading these softwares from the links above.

3 Tentative course schedule

The following is a tentative schedule of topics. Reading for R4DS refers to the numbering of the web version of the book. We may deviate from this schedule, so pay attention to announcements.

Week	Topic	Due	Reading
1	Introduction to R, RStudio and R Markdown		R4DS 1.4–1.6, 4, 27.1–27.4.3
2	Data visualization		R4DS 3, 28.2
3	**No class**	HW1	
4	Data types and representation	HW2	HoPR 1, 3-5, R4DS 20
5	Data frames and data manipulation	HW3	HoPR 3-4, R4DS 5.1–5.5, 10
6	Data summarization	HW4	R4DS 5.6–5.7
7	Data import and workflows	HW5	R4DS 6,8, 11.1–11.2, 11.6
8	**Exam**		
9	Tidy data and relational data		R4DS 12–13
10	**No class**		
11	Functions and control flow	HW6, Project proposal	R4DS 19, HoPR 6–7
12	Strings and factors	HW7	R4DS 14-15
13	Iteration and functional programming	HW8	R4DS 21, HoPR 9
14	Dates and times	HW9	R4DS 16
15	Working with many models and list-columns	HW10	R4DS 25
16	Debugging and performance enhancement		HoPR 10,E
		Project presentation	

4 Coursework & grading

There will be homework, a midterm exam, and a class project. Grading will be based on the following components:

- 50% Homework (10 assignments, lowest score dropped)
- 25% Exam (in-class)
- 25% Project

Learning to compute and program requires practice. So homeworks will be assigned on a weekly basis. These will mainly consist of exercises designed to reinforce the concepts covered in class during the previous week. Late homework will not be accepted, however you will be allowed to drop one homework score from your grade. I recommend completing all of the homeworks, even if you plan to drop one.

There will be one in-class exam. It will be open book/internet access, but absolutely no communicating with other humans will be allowed. The format of the exam will consist of conceptual questions as well as some computing and programming problems involving data.

Students will work in small groups on a final project consisting of producing an interactive Shiny app. The purpose of this project is for students to demonstrate the concepts and skills that they have acquired in this course, and to learn how to learn a new computing technology in a controlled environment. I will provide a list of topics. Each group will cooperate in the design, development, and making a presentation on the project.

5 Academic misconduct

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

6 Disability services

The University strives to make all learning experiences as accessible as possible. If you anticipate or experience academic barriers based on your disability (including mental health, chronic or temporary medical conditions), please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. SLDS contact information: slds@osu.edu; 614-292-3307; slds.osu.edu; 098 Baker Hall, 113 W. 12th Avenue.

Introduction to R for data science

Course proposal

2019-03-06

1 Rationale

R is a freely-available programming language and statistical computing environment. Over the last two decades, it has steadily gained in popularity and now serves as a major workhorse in modern statistical science research and practice. The vast majority of academic research in statistics (as well as other fields) is done in conjunction with a computing component which uses R as a programming tool. Recently, R has seen widespread adoption in other academic disciplines as well as in various industry settings. Nationwide Insurance, for example, is adopting R as a main programing language used in its analytics research. It is thus imperative that graduates of Ohio State University, aiming to continue their careers either through a graduate program, or by joining the data analytics work force, have a thorough knowledge of R.

2 Course objectives

The primary goal of the course is to teach students how to use R for data science by teaching students skills and underlying concepts of the R programming language and R package ecosystem for manipulation, visualization, and modeling of data, and for communicating the results of and enabling replication of their analyses.

The expected learning objectives for the students include:

- Understanding the basic concepts of the R programming language: expressions, objects, and data types;
- Understanding a grammar of graphics and being able to construct visualizations of data;
- Being able to efficiently manipulate and organize data;
- Being able to apply different numerical techniques for exploring and summarizing data;
- Understanding how data is represented, how to import data, the problems that may arise when importing data, and how to handle those problems;
- Understanding different programming paradigms and abstractions in R (for example, piping, iteration, and functional programming), and being able to recognize where and when these concepts can be applied;

- Being able to use different technologies for dynamically documenting, communicating the results of, and enabling replication of their data analyses with R and R Markdown.

3 Relationship to other courses/curricula

The proposed course has numerous connections to other courses in the Department of Statistics, as well as to courses in other departments. R is used in virtually every Statistics course that has a data analysis component, for example (this is not an exhaustive list)

- Undergraduate level: STAT 3201, 3202, 3301, 3302, 3303, 4302, 4620
- Graduate level: STAT 5301, 5302, 5550, 6410, 6450, 6500, 6520, 6530, 6550, 6560, 6570, 6620, 6730, etc. . .

However, none of these courses provide a systematic and comprehensive overview of R. Instead, existing courses demonstrate the use of R within the context of specific areas of applied statistics. Fundamental topics (e.g., data representations and manipulation, programing paradigms and abstractions, and performance limitations) and powerful features of R (e.g., creation of dynamic reports and websites) are not covered in existing courses. STAT 5730 will fill this significant gap in the curriculum, offering students the opportunity to gain foundational skills in R programing and an appreciation of the full potential of the R statistical computing environment.

The Department of Statistics is a partner in the newly created "Data Analytics" major at OSU. As part of the data analytics curriculum, the majors are required to take several elective courses—varying by specialization (biomedical informatics, business analytics and computational analytics). The proposed R course could serve as an elective course for either specialization. Our department is also home to a Masters of Applied Statistics. The graduate students enrolled in this program are all required to use the R language throughout their program. Adding this course as an elective to the MAS program would tremendously help both the students and the teachers in two ways: (i) it creates a uniform level of R programming knowledge among the students and (ii) it eliminates the need to teach R in other data analysis courses, thus freeing time for additional topics.

A large number of undergraduate and graduate students pursuing non-Statistics degrees (including students from the College of Engineering; the College of Food, Agriculture, and Environmental Science; etc.) are currently taking data analysis courses offered by the Department of Statistics (3301-3302 and 5301-5302). These students would benefit greatly from the proposed course as it complements the training they are receiving in these existing courses by providing more in-depth discussion of manipulating, summarizing, and analyzing data in R. Some of these students are currently completing the undergraduate or graduate minors in Statistics or the graduate minor in Statistical Data Analysis, and the proposed course could serve as an elective for these degrees.

Re: Concurrence for STAT 5730 Introduction to R for Data Science

Sivilotti, Paul

Mon 4/1/2019 9:44 AM

To: Lee, Yoonkyung <yklee@stat.osu.edu>;

Hi Yoon--

Our curriculum committee reviewed the proposal. CSE concurs with the course proposal.

Best wishes,
--paul

> On Mar 20, 2019, at 10:51 PM, Lee, Yoonkyung <yklee@stat.osu.edu> wrote:
>
> Dear Paul,
>
> We would like to request your concurrence on a new course, STAT 5730 Introduction to R for Data Science. This course is designed for our undergraduate and graduate students as an elective, teaching them skills and underlying concepts of the R programming language and computing environment for data analysis. Attached please find the sample syllabus and our rationale for the course proposal.
>
> Please let me know if you have any questions. We would appreciate getting your response within two weeks.
>
> Thank you!
>
> Yoon
>
> --
> Yoonkyung Lee
> Professor of Statistics
> Professor of Computer Science and Engineering (by courtesy)
> The Ohio State University
>
> <proposal.pdf><syllabus.pdf>

--

Prof. Paul A. G. Sivilotti Computer Science and Engineering
The Ohio State University 2015 Neil Ave., Columbus OH, 43210
614.292.5835, Fax 292.2911 http://www.cse.ohio-state.edu/~paolo