Course Syllabus

DS 636: Data Analytics with R Programming
Spring 2023

Instructor: Daming(David) Li, email: dli@njit.edu

Course Description and format:

This course will teach the hands-on skills in data analytics within the context of R but not limited to R, so that the skills we learn in the course are language neutral and can be used with any technologies/languages. The students will learn how to install and configure R necessary for an analytics programming environment and gain basic analytic skills via this high-level analytical language. The course covers fundamental knowledge in R programming. Popular R packages for data science will be introduced as working examples. The format of the course will include lectures by the instructor, lab exercises, class discussion, directed reading, and student presentations/projects. The exact format will depend on the size of enrollment and student background and will adjust according to the progress.

Goal

- We use R to teach this class but the content is for generic data science
- Focus on the skills that can be transferred to Python
- Familiarize you with the commonly used analytical techniques in Data Science
- Develop the way of data science thinking
- Learn how to preprocess, explore and interpret real data
- Learn how to model real problems using computational techniques

Prerequisite: Some basic knowledge of programming, probability and statistics. If in doubt about the prerequisites, please consult with the instructor for permission to take the class.

Attendance: You are supposed to attend all the classes. Participation is highly encouraged to make the class more interactive. In general, students who attend class regularly perform much better than those who come only occasionally. If you miss one class be sure to watch the recorded video and get notes, exercises, assignments, deadlines and announcements. If you are in the asynchronous online section, you will have access to the videos and watch at your convenience.

Textbooks (helpful but not required):

- R Programming for Data Science, by Roger D. Peng, https://leanpub.com/rprogramming
- Using R for Introductory Statistics, by John Verzani, Chapman & Hall/CRC, 2004, ISBN 1584884509
- Advanced R, by Hadley Wickham, ISBN 9781466586963.

Collaboration and Honor Code: Students may discuss problems together but must write up their own solutions. When writing up the solutions, students should write the names of people, if any, with whom they discussed the assignment. Note in particular that copying homework or programming assignments, in full or in part is forbidden. Students found cheating or plagiarizing will be immediately referred to the Dean of Students and the NJIT Committee on Professional Conduct and subject to Disciplinary Probation, a permanent marking on the record, possible dismissal, and an “F” grade in the course. All submitted assignments will be checked for similarities, and plagiarism and guilty students identified.

Grading:

The requirements of this course will consist of participating in lectures, homework, in class computing lab assignments, two exams and a project. The grading breakdown is the following:

- Homework, computing lab exercise (10%)
- Quiz (20%)
- Term Project (20%)
- Midterm (20%)
- Exam (30%)

Homework (10 %)

- Only use R in homework
- Try to do it independently, discussions allowed, but copying is forbidden.
- 25% penalization per late day;
- Not accepted more than 3 days late

Lab exercise

- Have a lab session every week
- Focus on R computing exercises
- We will solve some simple problems
- Post your answers by replying on canvas
- Some answers may be selected for discussion by the end of lab session.
- Some problems may become part of homework

Quiz (20%)

- Focus on course materials.
- 4 Quizzes
- Every other week
- Only R is allowed

Two Term Projects (20%)

- You can choose R or Python for your projects
- Use Jupyter
- Submit code and report to summarize what you have done and results you obtained.
- Prepare for presentation and demo.
- 1~4 students a group.
- More details to be announced on canvas
- Cheating/Copying is strictly prohibited.

Two Exams (50%)

- One midterm and one Final (20%+30%)
- In-class
- Final is cumulative
- Only R

Tentative course topics (Subject to changes according to progress)

1. Class overview and R basics
2. Advanced Data structures, IO & Control
3. Functions and Functional Programming
4. Manipulate Dataframe
5. Cloud Computing on Big Data
6. Interactive Data Visualization
7. Graph Theory and Analytics
8. Probability and Statistics for Data Science
9. Text Analytics, NLP & Similarity
10. Data Clustering
11. Linear models and metrics.
12. Feature and Model selection and SVM
13. Trees, kNN, NN and Final Review