**CAP 3321**

**RI: Introduction to Data Science**

Instructor: Dr. Lianfen Qian

Office: SE 244

Office Hours: TBA

Email: Lqian@fau.edu

**Catalog description**: This research-intensive course will survey the foundational topics in data science. It covers the following topics: Data acquisition, data manipulation, data exploration and visualization, data analysis with statistics and machine learning, data at scale via working with big data. The course will use statistical software to work through real-world examples that illustrate these concepts. Concurrently, students learn some statistical and mathematical foundations that power the data scientific approach to problem solving.

**Textbook:**  Data Science A-ZTM: Real-life data science exercises included by Kirill Eremenko/ UDemy

**Course description:** Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

This course address undergraduate research and inquiry components on (1) knowledge, (2) formulating questions, (3) plan of action, (4) critical thinking, (5) ethical conduct, and (6) communication.

**Course objectives**: Students who successful completes the course will grasp the coverage of data science process and usage of common data science tools. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection, processing and integration, data visualization, descriptive modeling, predictive modeling, data product creation and evaluation, and effective communication. Students will have hands-on experience with statistical inference and data mining software for research projects.

**Research Intensive Designation:** This course contains an assignment or multiple assignments designed to help students conduct research and inquiry at an intensive level.  If this class is selected to participate in the university-wide assessment program, students will be asked to complete a consent form and submit electronically some of their research assignments for review.  Visit the Office of Undergraduate Research and Inquiry (OURI) for additional opportunities and information at <http://www.fau.edu/ouri>.

The URI portion of the course will address all six Student Learning Objectives:

1. **Knowledge**: Common base of knowledge required for effective data preprocessing, data visualization, data learning, model selection and reproducible research. Students will grasp a set of key skills on statistical learning and computing in data analytics. Students will also show knowledge of tools and practical skills needed to preprocessing and managing data from various sources for both structured and unstructured massive data.
2. **Formulation of Questions**: Students are required to develop research statement in which they specifically address their research questions for real data sets or simulation study. Students are expected to formulate their research questions into subject related hypotheses which are clear and concise to the research problem, ready to be tested or answered through statistical modeling and simulation. When appropriate, the students should be able to break down principal problems into smaller solvable sub-problems.
3. **Plan of Action**: Students will create a plan of action for term projects of research intensive courses in the program such as for Introduction to Data Science and Computational Statistics that will encompass the following elements: (i) scope of the study; (ii) literature review; (iii) planning context; (iv) problem statement and research methodology; (v) analysis and report findings; (vi) presenting the results to general audience. The students will develop hypothesis if needed, identify methods of analysis and select appropriate statistical techniques. Using the course timeline as a template, each student is expected to develop her/his own project management plan with specific tasks related to the topic in consideration.
4. **Critical Thinking:** Students will demonstrate critical thinking skills by formulating research questions, applying appropriate selection criteria for real data analysis or simulation study, taking into consideration multiple perspectives, and examining implications and consequences of an action or planning alternative.
5. **Ethical conduct:** All students are required to familiarize themselves with the rules of academic integrity. Student projects involving primary data collection through website will be credited in their term written research projects. Students are required to be loyal to the original data with confidential information removed.
6. **Communication:** Students will be required to write and present their project reports professionally. They are required to submit research report (e.g., analysis, findings and recommendations), and develop a webpage uploaded to rpubs.com for instant to communicate research results as outlined in SLO-3. Students are expected to demonstrate knowledge of writing technical report and orally presenting their findings. Advanced visualization techniques are also required for students to incorporate research findings in planning documents and present them through a real data project.

**Reference books:**

1. Introduction to data science by Jeffrey Stanton, Syracuse University, <https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf>.
2. Data science from scratch by Joel Grus <http://proquest.safaribooksonline.com/book/databases/9781491901410>
3. [https://en.wikibooks.org/wiki/Data\_Science:\_An\_Introduction](https://en.wikibooks.org/wiki/Data_Science%3A_An_Introduction)
4. <http://www.cs.tau.ac.il/~apartzin/ds2015/DS_CourseIntro.pdf>
5. <https://www2.stat.duke.edu/courses/Fall15/sta112.01/>
6. Reference course at Berkeley: <https://bcourses.berkeley.edu/courses/1377158/pages/cs-194-16-introduction-to-data-science-fall-2015>
7. <https://bcourses.berkeley.edu/courses/1377158>
8. https://medium.freecodecamp.com/i-ranked-all-the-best-data-science-intro-courses-based-on-thousands-of-data-points-db5dc7e3eb8e#.1kz3fwrm1

**Prerequisites:** MAD 2502 or COP 2220 with minimum grade C

**Grading policy**:

Class attendance and participation: 10%

Biweekly homework/lab assignment: 40%

Midterm project: 20%

Final project: 30%

Grading scale: A/A-: 90-100%, B+/B/B-: 80-89%, C+/C: 70-79%, D: 60-69%, F: <59%.

**Course outline:**

Week 1: Introduction to R/Rstudio and set up git/GitHub

Week 2: R/Rstudio/GitHub demo, introduce to data visualization

Week 3: Hands-on data visualization, data wrangling via real data

Week 4: Data types, data expeditions, the language of models

Week 5: Transforming data, interactions and model selection

Week 6: Modeling recap and wrap up real data expedition, prediction and model validation

Week 7: Work on midterm project and midterm project presentations

Week 8: Hypothesis testing via simulation

Week 9: Functional hypothesis testing

Week 10: Testing for independence and applications

Week 11: Central limit theorem based inference

Week 12: Interactive visualizations with Shiny

Week 13: Scraping data off the web and hands-on analysis

Week 14: Intro to Bayesian inference

Week 15: Final project presentations

**Description of the RI assignments:**

**Assignment 1 (Data Preprocessing and Management)**: In week 4 of the course, students will prepare and clean a real data for analysis and publish the computing R scripts to a repository website. Students will submit the first tech report on the data preprocessing and preliminary exploratory data analysis. This submission will be limited to five page preliminary tech report which will include a description of the real data, statistical thinking and formulation of the problem, a description of the methods used to preprocessing the data, a summary of the findings and a conclusion.

**Assignment 2 (Midterm Project)**: In week 8 of the course, students will prepare a tech report and publish the computing R scripts to a repository website. Students submit the second tech report on the data preprocessing and intermediate level data analysis including method of explorations and model selection. This submission will be limited to eight-page tech report which will include a description of the real data, statistical thinking and formulation of the problem, data preprocessing, a description of the methods considered and model selection criteria, a summary of the findings with performance measure and a conclusion.

**Assignment 3 (Final Project):** In weeks 9-14 of the course, students will write a 10 page tech report following tech report guideline on data management, methods of explorations from simple to advanced modeling techniques, model selection criteria, and publish the computing R scripts to a repository website. Students will present their analysis and results to classmates and submit the final tech report by the end of the semester. This submission will be limited to 10-15 page tech report/scientific paper which will include a description of the real data, statistical thinking and formulation of the problem, data preprocessing and management procedure, a description of the methods considered and model selection criteria, a summary of the findings with performance measure, comments on complication and limitations and conclusions.

**Incomplete grades**

A grade of I (incomplete) will only be given under certain conditions and in accordance with the academic policies and regulations put forward in FAU's University Catalog. The student has to show exceptional circumstances why requirements cannot be met. A request for an incomplete grade has to be made in writing with supporting documentation, where appropriate.

**Classroom etiquette policy**

University policy on the use of electronic devices states: “In order to enhance and maintain a productive atmosphere for education, personal communication devices, such as cellular telephones and pagers, are to be disabled in class sessions.”

**Disability policy statement**

In compliance with the Americans with Disabilities Act (ADA), students who require reasonable accommodations due to a disability to properly execute coursework must register with Student Accessibility Service (SAS) ---in Boca Raton, SU 133 (561-297-3880) ; in Davie, LA 203 (954-236-1222); or in Jupiter, SR 110 (561-799-8585)---and follow all SAS procedures.

**Academic integrity**

Students at Florida Atlantic University are expected to maintain the highest ethical standards. Academic dishonesty, including cheating and plagiarism, is considered a serious breach of these ethical standards, because it interferes with the University mission to provide a high quality education in which no student enjoys an unfair advantage over any other. Academic dishonesty is also destructive of the University community, which is grounded in a system of mutual trust and places high value on personal integrity and individual responsibility. Harsh penalties are associated with academic dishonesty. For more information, see <https://www.fau.edu/ctl/4.001_Code_of_Academic_Integrity.pdf>