**R25 Emerging Scholars 2024 Summer Institute --- Week 2 (May 20-24)**

**Introduction to Big Data Analytics**

**Location: Discovery Building (see timetable for locations)**

**Instructor(s):** Yanan Zhang, PhD Andy Ortaglia, PhD   
**Email:** [yanan@email.sc.edu](mailto:yanan@email.sc.edu) [ortaglia@mailbox.sc.edu](mailto:ortaglia@mailbox.sc.edu)   
**Phone:** 803-777-7563 803-777-7492

**Course Description:** The “Introduction to Big Data Analytics” session is part of a six-week summer training program funded by NIH. This one-week session will cover topics of descriptive statistics, hypothesis testing, regression models, clustering (unsupervised learning method), tree analyses (supervised learning method), and techniques for handling missing values. Through lectures and hands-on labs, this session aims to help students understand these statistical topics and enable students to gain practical experiences in the field of data analysis.

**Course Objectives:** A student who successfully completes this program will:

* improve/develop an understanding of descriptive statistics, probability, sampling distributions and hypothesis testing
* become familiar with traditional regression techniques such as linear regression, ANOVA and generalized linear models
* become familiar with pattern analysis techniques such as clustering and tree analysis.
* develop an understanding of the fundamentals of imputation methods for handling missing data
* gain experience using the ‘R’ environment for statistical computing
* be able to properly interpret statistical results to address a scientific research question.

**Course Requirement (including presentation, homework and lab assignment):**

Each day of the short course adheres to a structured format aimed at enhancing learning and fostering engagement through a combination of lectures, lab sessions, and group discussions. During the lecture sessions, instructors give presentations encompassing both theoretical concepts and practical applications pertinent to the day's subject. Following the lectures, students participate in lab sessions where they acquire hands-on experience with the content covered earlier. Here, they actively engage with statistical software R, under the guidance of instructors. Subsequently, both students and instructors participate in a discussion related to the day's topic, for example, interpreting analysis results from the lab session, addressing challenges encountered, or exploring alternative analysis methods.

To assess understanding and reinforce learning objectives, quick quizzes are administered (See Evaluation Criteria part for details). This structured approach ensures that students integrate theoretical knowledge with practical skill development and collaborative learning experiences.

**Course Materials/References (any downloads can be found in the secure portal at** [**https://bigdata.sc.edu/r25-e-scholar-account/**](https://bigdata.sc.edu/r25-e-scholar-account/)**):**

Materials including handouts, slides, and R codes will be provided.

**Evaluation Criteria:**

To pass the course, students must demonstrate a satisfactory level of understanding and application of statistical concepts as evidenced by their performances, including attendance, quizzes, and participation in group discussions. A quiz related to the material covered the previous day will be administered at the start of each subsequent day. For the contents covered in the last day, a quiz will be given at the end session of the lecture.

Grading:

|  |  |  |
| --- | --- | --- |
| Criteria | Weight (%) | Details |
| Attendance | 25 | Each day accounts for 5%. |
| Quizzes | 25 | Five quizzes, each accounting for 5%. |
| Group Discussion | 50 | Active participation in group discussions. |

Pass/Fail Threshold:

A total grade of 80 and above is considered a pass. Note: The attendance, quizzes, and group discussion criteria collectively contribute to the overall evaluation. Students must achieve the pass threshold to successfully complete the course.

**Workshop Schedule and Activities**

|  |  |  |
| --- | --- | --- |
| **Day 1: Monday, May 20 Basic Statistics and Hypothesis Testing** | | |
| **Time and location** | **Topic** | **Activity** |
| 8:30-10:00 DISC 331 | -numerical summary measures (location, variation, and shape)  -probability distributions and sampling distributions  -confidence intervals | Lecture I |
| 10:15-11:45 DISC 431 | -introduce R software  -import data and prepare data  -conduct descriptive statistics  -create tables and plots of the data | Lab I |
| Lunch on your own | | |
| 1:15-2:15 DISC 331 | -introduce some of the large public health data sets  -discuss the types of research/scientific questions can be addressed | Group Discussion |
| 2:30-4:00 DISC 331 | -hypothesis testing  -inference on means and proportions | Lecture II |
| 4:15-5:30 DISC 431 | -conduct hypothesis tests  -interpret results and make conclusion | Lab II |
| **Day 2: Tuesday, May 21  Regression Analysis** | | |
| 8:30-10:00 DISC 331 | -ANOVA, simple linear regression, multiple linear regression  -confounding and interaction | Lecture I |
| 10:15-11:45 DISC 431 | -conduct linear regressions in R -do model diagnostics  -interpret results and make conclusion | Lab I |
| Lunch on your own | | |
| 1:15-2:15 DISC 431 | -present results from Lab I  -discuss alternative approaches when model assumptions not satisfied | Group Discussion |
| 2:30-4:00 DISC 331 | -measures of association (e.g., odds ratio, relative risk, incidence rate ratio)  -Logistic regression  -Poisson regression | Lecture II |
| 4:15-5:30 DISC 431 | -conduct Logistic regression and Poisson regression in R  -interpret results and make conclusion | Lab II |
| **Day 3: Wednesday, May 22  Clustering** | | |
| 8:30-10:00 DISC 302 | -Clustering distance measures  -K-means algorithm | Lecture I |
| 10:15-11:45 DISC 431 | -introduce R packages  -compute K-means clustering in R  -interpret results and make conclusion | Lab I |
| Lunch on your own | | |
| 1:15-2:15 DISC 431 | -present results from Lab I | Group Discussion |
| 2:30-4:00 DISC 259 | -Hierarchical agglomerative clustering algorithm -dendrograms and heatmap  -cluster validity | Lecture II |
| 4:15-5:30 DISC 431 | -introduce R packages  -conduct hierarchical agglomerative clustering in R  -create dendrograms and heatmap  -interpret results and make conclusion | Lab II |
| **Day 4: Thursday, May 23 Tree Analysis** | | |
| 8:30-10:00 DISC 331 | -decision trees: build a tree, select the best split, overfitting  -classification performance evaluation | Lecture I |
| 10:15-11:45 DISC 431 | -introduce R packages  -conduct regression tree in R  -conduct classification tree in R -interpret results and make conclusion | Lab I |
| Lunch on your own | | |
| 1:15-2:15 DISC 431 | -present results from Lab I | Group Discussion |
| 2:30-4:00 DISC 331 | -random forest | Lecture II |
| 4:15-5:30 DISC 431 | -introduce R packages  -conduct random forest in R  -interpret results and make conclusion | Lab II |
| **Day 5: Friday, May 24 Handling Missing Data** | | |
| 8:30-10:00 DISC 140 | -missing data mechanism  -identify/explore missing values patterns  -impact of missing data | Lecture I |
| 10:15-11:45  DISC 431 | -introduce R packages  -identification and visualization of missing data  -deal with missing data in R | Lab I |
| Lunch on your own | | |
| 1:15-2:15 DISC 431 | -present results from Lab I  -discuss main types of missing data | Group Discussion |
| 2:30-4:00 DISC 140 | -approaches to deal with missing data: complete-cases analysis, available data analysis, imputation and weighting methods | Lecture II |
| 4:15-5:30 DISC 431 | -deal with missing data in R | Lab II |

**University Policies**

|  |  |
| --- | --- |
| **Incomplete coursework** | Incomplete coursework is a major inconvenience for students and instructors. I expect you to do everything in your power to avoid this situation. Legitimate excuses include verified illnesses and family emergencies. No incompletes will be granted only in accordance with [university policy](https://sc.edu/about/offices_and_divisions/advising/withdrawal/dropping_courses/). |
| **Classroom conduct** | I will conduct this class in an atmosphere of mutual respect. I encourage your active participation in class discussions. Each of us may have strongly differing opinions on the various topics of class discussions. The conflict of ideas is encouraged and welcome. The orderly questioning of the ideas of others, including mine, is similarly welcome. However, I will exercise my responsibility to manage the discussions so that ideas and argument can proceed in an orderly fashion. You should expect that if your conduct during class discussions seriously disrupts the atmosphere of mutual respect I expect in this class, I can’t help but ask you to stop participating. |
| **Academic misconduct** | You are expected to do your own academic work and cite sources as necessary. Students are expected to adhere to all requirements of the Carolinian Creed ([www.sa.sc.edu/creed/](http://www.sa.sc.edu/creed/)). Please especially note: as a member of this training program, you are held accountable to this Creed even if you violate it inadvertently or are not a registered USC student. Any episode of dishonesty, cheating, or plagiarism in any form is cause for failure of an assignment, an examination, or a course. Students may want to refresh their understanding of the appropriate use of citations when drafting papers and other assignments to prevent inadvertent plagiarism stemming from lack of information. In addition, program leadership may decide to withdraw their support of your participation in this training program. |
| **Sexual harassment** | "Sexual harassment" means unwelcome sexual advances, requests for sexual favors, and/or other verbal or physical conduct of a sexual nature. Such conduct has the purpose or effect of unreasonably interfering with an individual's work or academic performance or creating an intimidating, hostile, or offensive working or academic environment in any University activity or program. Such behavior is not acceptable in the University setting. For additional information, please consult [Equal Opportunity Programs Policy](https://www.sc.edu/about/offices_and_divisions/equal_opportunities_programs/documents/sexual_harassment.pdf). |
| **Accessibility, Disability, and Triggers** | The USC Arnold School of Public Health is committed to providing equitable access to learning opportunities for all students.  If you have, or think you may have, a disability (e.g., mental health, attentional, learning, chronic health, sensory, or physical), please notify me so confidential discussion regarding equitable access and reasonable accommodations can take place. |
| **Diversity** | The University provides equal access to and opportunity in its programs and facilities, without regard to race, color, creed, religion, national origin, gender, age, marital status, disability, public assistance status, veteran status, sexual orientation, gender identity, or gender expression. For more information, please consult the [Student Non-Discrimination and Non-Harassment Policy](https://www.sc.edu/about/offices_and_divisions/equal_opportunities_programs/documents/student_non-discrimination_and_non-harassment_policy.pdf). |
| **Title IX and Gendered Pronouns** | This course affirms equality and respect for all gendered identities and expressions. Please don’t hesitate to correct the instructor regarding your preferred gender pronoun and/or name if different from what is indicated on the official class roster. Likewise, the instructor is committed to nurturing an environment free from discrimination and harassment. Consistent with Title IX policy, please be aware that the instructor is obligated to report information that students provide about a situation involving sexual harassment or assault. |