2020

Introduction to Data Science

North Carolina Agricultural and Technical State University

NORTH CAROLINA AGRICULTURAL AND TECHNICAL STATE UNIVERSITY

COURSE SYLLABUS

College Name: College of Science and Technology
Department Name: Department of Mathematics & Statistics
Course Name: Introduction to Data Science

COURSE INFORMATION

- Course Number/Section: STAT 707
- Term:
- Semester Credit Hours: 3
- Times and Days:
- Class Location:

INSTRUCTOR CONTACT INFORMATION

- Instructor:
- Office Location:
- Office Phone:
- Email Address:

Faculty must notify students of the approximate time and method they can expect to receive an answer to all communications (e.g., email, phone, course messages). Excluding holidays, the response should be provided within 48 hours.

If there's a graduate teaching assistant assigned to work with this course, please include their names also.

STUDENT HOURS

These are times students may visit the professor without an appointment to request the assistance they need.

NOTE: Students are responsible for reading, understanding, and following the syllabus.

: AM ☐ / PM ☐ – : AM ☐ / PM ☐

Monday ☐ Tuesday ☐ Wednesday ☐ Thursday ☐ Friday ☐

COURSE PREREQUISITES

Graduate Standing and permission of the instructor.

COURSE DESCRIPTION

This course covers the data science concepts and skills needed to collect, manipulate, compute with, and interpret data. The students will learn to write R, SAS and/or Python programs that manipulate and visualize data, and they will learn statistical inference, predictive modeling and machine learning techniques essential to analyzing and presenting the data

STUDENT LEARNING OBJECTIVES/OUTCOMES (SLO)

Learning outcomes should be specific, measurable, and focused on the content knowledge the students are expected to master and not what the faculty will teach.

If the course is a General Education Course, the SLO should be listed and labeled as "General Education."

SLO 1: Recognize the main tasks a data scientist would typically work on.
SLO 2: Develop a strong understanding of the programming environment R as used to perform the data science tasks
SLO 3: Recognize the importance of data preparation and data wrangling in the data analysis process, and use tools from the "tidyverse" package in R (and "Pandas" in Python) to perform these tasks.
SLO 4: Perform data visualizations using the "ggplot2" package in R (and the "matplotlib" and "seaborn" libraries in Python).
SLO 5: Understand the foundations of statistical inference and predictive modeling techniques.
SLO 6: Build and interpret different state-of-the-art predictive models, such as Linear/Logistic Regression, K-nearest Neighbors, Classification/Regression Trees, and Random Forests.

REQUIRED TEXTBOOKS AND MATERIALS

Any course-level subscriptions and tools linked in Blackboard Learn learning management system (LMS) should be listed here. The Blackboard LMS must have links to their student data privacy statement.

REQUIRED TEXTS:

Irizarry, R. A. (2020). Introduction to data science: Data analysis and prediction algorithms with R. Chapman and Hall/CRC.

REQUIRED MATERIALS:

Statistical software: R and Python

SUGGESTED COURSE MATERIALS

SUGGESTED READINGS/TEXTS:

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize and model data*. Beijing: O'Reilly.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. New York: Springer.

Xie, Y., Allaire, J. J., & Grolemund, G. (2019). *R Markdown: The definitive guide*. Boca Raton: CRC Press, Taylor & Francis Group.

SUGGESTED MATERIALS:

GRADING POLICY

ASSIGNMENTS AND GRADING POLICY

94% and above	A		76% - 74%	C
93% - 90%	A-		73% - 70%	C-
89% - 87%	B+		69% - 67%	D+
86% - 84%	B		66% - 64%	D
83% - 80%	B-		63% - 0%	F
79% - 77%	C+			

For GRADUATE COURSES: See 2019-2020 Graduate Catalog p.38 for graduate grading scale and Non-Graded Courses

GRADING ALLOCATION

Course grades are based on a weighted grading scale of 100%. The breakdown for the course is as follows: *[Faculty, please adjust according to your course.]*

Category	# of Activities	Percentage Grade Weight
Discussion Boards	7	5%
Homeworks	15	30%
Quizzes	8	5%
Data Analysis Projects	3	9%
Data Analysis Project Final	1	6%
Midterm Exam	1	20%
Final Exam	1	25%
Total	**34**	**100%**

COURSE POLICIES

USE OF BLACKBOARD AS THE LEARNING MANAGEMENT SYSTEM

Blackboard is the primary online instructional and course communications platform. Students can access the course syllabus, assignments, grades, and learner support resources. Students are encouraged to protect their login credentials, complete a Blackboard orientation, and log in daily to the course.

Note: Uploading assignments through Blackboard presents a challenge for Chromebook users in locating the files for submission. If you use a Chromebook, please be sure you also have access to a Mac computer or Windows computer so you can fully participate in your Blackboard class. For more information about student computer recommendations, please visit https://hub.ncat.edu/administration/its/computer-recommendations.php.

MAKE-UP EXAMS

See << Update Academic Year >>*Undergraduate Bulletin:*

https://www.ncat.edu/provost/academic-affairs/bulletins/index.php

For GRADUATE STUDENTS: See 2019-20 Graduate Catalog p. 54

EXTRA CREDIT

LATE WORK

SPECIAL ASSIGNMENTS

For GRADUATE STUDENTS: FAILING TO MEET COURSE REQUIREMENTS (Graduate Catalog p.40)

For GRADUATE STUDENTS: CLASS ATTENDANCE (see 2019-20 Graduate Catalog p. 53-54)

Students are expected to attend class and participate on a regular basis in order to successfully achieve course learning outcomes and meet federal financial aid requirements (34 CFR 668.22). Class attendance in online courses is defined as active participation in academically-related course activities. Active participation may consist of course interactions with the content, classmates, and/or the instructor. Examples of academically-related course activities include, but are not limited to:

- Completing and submitting assignments, quizzes, exams, and other activities within Blackboard or through Blackboard (3rd-party products).
- Participating in course-related synchronous online chats, discussions, or meeting platforms such as Blackboard Collaborate in which participation is tracked.

CLASSROOM CITIZENSHIP

Courtesy, civility, and respect must be the hallmark of your interactions.

COMPLIANCE WITH THE AMERICANS WITH DISABILITIES ACT

North Carolina A&T State University is committed to following the requirements of the Americans with Disabilities Act Amendments Act (ADAAA) and Section 504 of the Rehabilitation Act.
If you need an academic accommodation based on the impact of a disability, you must initiate the request with the Office of Accessibility Resources (OARS) and provide documentation in accordance with the Documentation Guidelines at N.C. A&T. Once documentation is received, it

will be reviewed. Once approved, you must attend a comprehensive meeting to receive appropriate and reasonable accommodations. If you are a student registered with OARS, you must complete the Accommodation Request Form to have accommodations sent to faculty.

OARS is located in Murphy Hall, Suite 01 and can be reached at 336-334-7765, or by email at accessibilityresources@ncat.edu. Additional information and forms can be found on the internet at https://www.ncat.edu/provost/academic-affairs/accessibility-resources/index.php.

Please note: Accommodations are not retroactive and begin once the Disability Verification Form is provided to faculty.

TITLE IX

North Carolina A&T State University is committed to providing a safe learning environment for all students—free of all forms of discrimination and harassment. Sexual misconduct and relationship violence in any form are inconsistent with the university's mission and core values, violates university policies, and may also violate federal and state law. Faculty members are considered "Responsible Employees" and are required to report incidents of sexual misconduct and relationship violence to the Title IX Coordinator. If you or someone you know has been impacted by sexual harassment, sexual assault, dating or domestic violence, or stalking, please visit the Title IX website to access information about university support and resources. If you would like to speak with someone confidentially, please contact Counseling Services at 336-334-7727 or the Student Health Center at 336-334-7880.

TECHNICAL SUPPORT

If you experience any problems with your A&T account, you may call Client Technology Services (formerly Aggie Tech Support and Help Desk) at 336-334-7195, or visit https://hub.ncat.edu/administration/its/dept/ats/index.php.

FIELD TRIP POLICIES / OFF-CAMPUS INSTRUCTION AND COURSE ACTIVITIES

If applicable:

Off-campus, out-of-state, foreign instruction, and activities are subject to state law and university policies and procedures regarding travel and risk-related activities. Information regarding these rules and regulations may be found at https://www.ncat.edu/campus-life/student-affairs/index.php.

STUDENT HANDBOOK

https://www.ncat.edu/campus-life/student-affairs/departments/dean-of-students/student-handbook.php

STUDENT TRAVEL PROCEDURES AND STUDENT TRAVEL ACTIVITY WAIVER

https://hub.ncat.edu/administration/student-affairs/staff-resources/studen_activity_travel_waiver.pdf

OTHER POLICIES *(e.g., Copyright Guidelines, Confidentiality, etc.)*

STUDENT HANDBOOK

https://www.ncat.edu/campus-life/student-affairs/departments/dean-of-students/student-handbook.php

Graduate Catalog

SEXUAL MISCONDUCT POLICY

https://www.ncat.edu/legal/title-ix/sexual-harassment-and-misconduct-policies/index.php

FAMILY EDUCATIONAL RIGHTS AND PRIVACY ACT (FERPA)

https://www.ncat.edu/registrar/ferpa.php

STUDENT COMPLAINT PROCEDURES

https://www.ncat.edu/current-students/student-complaint-form.php

STUDENT CONDUCT AND DISCIPLINE

North Carolina A&T State University has rules and regulations that govern student conduct and discipline meant to ensure the orderly and efficient conduct of the educational enterprise. It is the responsibility of each student to be knowledgeable about these rules and regulations.

Please consult the following about specific policies such as academic dishonesty, cell phones, change of grade, disability services, disruptive behavior, general class attendance, grade appeal, incomplete grades, make-up work, student grievance procedures, withdrawal, etc.:

- Undergraduate Bulletin
  https://www.ncat.edu/provost/academic-affairs/bulletins/index.php
- Graduate Catalog
  https://www.ncat.edu/tgc/graduate-catalog/index.php
- Student Handbook
  https://www.ncat.edu/campus-life/student-affairs/departments/dean-of-students/student-handbook.php

ACADEMIC DISHONESTY POLICY

Academic dishonesty includes but is not limited to the following:

1. Cheating or knowingly assisting another student in committing an act of cheating or other academic dishonesty;
2. Plagiarism (unauthorized use of another's words or ideas as one's own), which includes but is not limited to submitting exams, theses, reports, drawings, laboratory notes or other materials as one's own work when such work has been prepared by or copied from another person;
3. Unauthorized possession of exams or reserved library materials; destroying or hiding source, library or laboratory materials or experiments or any other similar actions;
4. Unauthorized changing of grades, or marking on an exam or in an instructor's grade book or such change of any grade record;
5. Aiding or abetting in the infraction of any of the provisions anticipated under the general standards of student conduct;
6. Hacking into a computer and gaining access to a test or answer key prior to the test being given. N.C. A&T reserves the right to search the emails and computers of any student suspected of such computer hacking (if a police report of the suspected hacking was submitted prior to the search); and
7. Assisting another student in violating any of the above rules.

A student who has committed an act of academic dishonesty has failed to meet a basic requirement of satisfactory academic performance. Thus, academic dishonesty is not only a basis for disciplinary action, but may also affect the evaluation of a student's level of performance. Any student who commits an act of academic dishonesty is subject to disciplinary action.

In instances where a student has clearly been identified as having committed an act of academic dishonesty, an instructor may take appropriate disciplinary action, including loss of credit for an assignment, exam, or project; or awarding a grade of "F" for the course, **subject to review and endorsement by the chairperson and dean**.

For GRADUATE STUDENTS: Reference for academic dishonesty – 2010-2020 Graduate Catalog, p.58-59

For GRADUATE STUDENTS: STUDENT RELIGIOUS OBSERVANCE (see Graduate Catalog, p.55)

ASSIGNMENTS AND ACADEMIC CALENDAR

Include topics, reading assignments, due dates, exam dates, withdrawal dates, pre-registration and registration dates, all holidays, and convocations.*

THE WEEK OF MM/DD/YY	SUBJECT	UNIT LEARNING OUTCOMES (ULO)	READING IN TEXT, ACTIVITY, HOMEWORK, EXAM
	Unit 1: Introduction & Getting Started	ULO 1: Describe the elements of the emerging field of data science and discuss its importance. (SLO 1) ULO 2: Recognize the qualifications and roles of data scientists and the career opportunities available to them. (SLO 1) ULO 3: Describe the elements of data science projects. (SLO 1) ULO 4: Describe the different types of data and their characteristics (3Vs: Volume, Variety, Velocity). (SLO 1) ULO 5: Recognize the tools in the data scientist's toolbox. ULO 6: Recall how and where to find help for data science projects. (SLO 1) ULO 7: I Demonstrate an understanding of R and RStudio in order to compare their capabilities as used for data science projects. (SLO 2)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 1:Getting Started with R and RStudio b. Chapter 38.1–38.2: Reproducible projects with RStudio and R markdown 2. **Complete**: Discussion Board #1 (N/A) 3. **Complete**: Quiz #1 (ULO 1-8) 4. **Complete**: Homework #1 (ULO 1-8)

		ULO 8: Install R packages from different repositories (CRAN, GitHub, and Bioconductor), create projects and R Markdown reports in RStudio. (SLO 2)	
	Unit 2: Basic Coding and Plotting in R	ULO 1: Perform basic coding operations in RStudio. (SLO 2) ULO 2: Understand the fundamental structure behind coding in R. (SLO 2) ULO 3: Recognize different types of plots that can be created in base R. (SLO 2) ULO 4: Recognize the limitations in the base R plotting functions. (SLO 2)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 2: R Basics 2. **Complete**: Discussion Board #2 (ULO 1-4) 3. **Complete**: Quiz #2 (ULO 1-4) **Complete**: Homework #2 (ULO 1-4)
	Unit 3: Basic Programming in R	ULO 1: Perform basic programming tasks in RStudio such as writing conditional expressions and defining their own functions. (SLO 2) ULO 2: Apply functions repeatedly using for-loops and functionals such as apply, lapply, and sapply. (SLO 2) ULO 3: Write vectorized functions. (SLO 2)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 3: Programming basics 2. **Complete**: Quiz #3 (ULO 1-3) 3. **Complete**: Homework #3 (ULO 1-3)
	Unit 4: Manipulating data frames with dplyr from the tidyverse collectio n	ULO 1: Explain the difference between tidy and untidy data. (SLO 2-3) ULO 2: Describe the tidyverse collection of R packages. (SLO 2-3) ULO 3: Perform various data manipulation tasks using functions from the dplyr package (including “mutate”, “filter”, “select” and, “summarize”, etc.). (SLO 2-3)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 4: The tidyverse 2. **Complete**: Discussion Board #3 (ULO 1-3) 3. **Complete**: Homework #4 (ULO 1-3) 4. **Complete**: Quiz #4 (ULO 1-3) 5. **Complete**: Data Analysis Project#1 (ULO 1-3)
	Unit 5: Importing data into R	ULO 1: Import data files of different kinds into R using functions from base-R. (SLO 2-3) ULO2: Import data files of different kinds into R using functions from the readr and readxl packages in tidyverse.	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 5: Importing data 2. **Complete**: Homework #5 (ULO 1-3)

		(SLO 2-3) ULO 3: Describe the differences between the base-R and the tidyverse data importing functions. (SLO 2-3)	
	Unit 6: Data visualization using ggplot2 - Part I	ULO 1: Recognize the role of data visualizations as tools for summarizing and discovering patterns in datasets. (SLO 4) ULO 2: Describe the components of a graph. (SLO 4) ULO 3: Describe the different components of a ggplot (i.e., aesthetics, geometries, layers, scales, labels, titles, and annotations). (SLO 4)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 6: Introduction to data visualization b. Chapter 7: ggplot2 2. **Complete**: Discussion Board #4 (ULO 1-3) 3. **Complete**: Quiz #5 (ULO 1-3) 4. **Complete**: Homework #6 (ULO 1-3) 5. **Complete**: Data Analysis Project #2 (ULO 1-3)
	Unit 7: Data visualization using ggplot2 - Part II	ULO 1: Describe the different ways to visualize and compare distributions. (SLO 4) ULO 2: Use the ggplot2 package to produce powerful visualizations for complex data. (SLO 4)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 8: Visualizing data distributions b. Chapter 9: Data visualization in practice 2. **Complete**: Quiz #6 (ULO 1-2) 3. **Complete**: Homework #7 (ULO 1-2)
	Unit 8: Data visualization using ggplot2 - Part III	ULO 1: Explain and implement the main data visualization principles (SLO 2-3) ULO 2: Describe robust tools for summarizing data (SLO 2-3)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 10: Data visualization principles b. Chapter 11: Robust summaries 2. **Complete**: Quiz #7 (ULO 1-2) 3. **Complete**: Homework #8 (ULO 1-2)
	Unit 9: Data wrangling: reshaping data with tidyr	ULO 1: Use functions from the tidyr package to make a raw dataset tidy. (SLO 1-3) ULO 2: Recognize the uses of the different tidyr functions: gather, spread, separate, and unite. (SLO 1-3)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 21: Reshaping Data 2. **Complete**: Discussion Board #5 (ULO 1-2) 3. **Complete**: Homework #9 (ULO 1-2) 4. **Complete**: Midterm Exam (ULO Unit 1-9)
	Unit 10: Data	ULO 1: Join multiple data	1. **Read Textbook**: Irizarry, R. A.

	wrangling: joining data tables with dplyr	tables using functions from the dplyr package. (SLO 1-3) ULO 2: Distinguish between the different types of dplyr joins (mutating joins and filtering joins). (SLO 1-3) ULO 3: Describe the difference between binding and set operation commands in base R and in dplyr. (SLO 1-3)	(2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 22: Joining tables 2. **Complete**: Discussion Board #6 (ULO 1-3) 3. **Complete**: Homework #10 (ULO 1-3) 4. **Complete**: Quiz #8 (ULO 1-3)
	Unit 11: Web scraping with rvest	ULO 1: Use the rvest package to scrape data from web pages. (SLO 1-3) ULO 2: Distinguish between the different kinds of html nodes and describe how to extract them from web pages using tools from the rvest tools. (SLO 1-3)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 23: Web Scraping 2. **Complete**: Discussion Board #7 (ULO 1-2) 3. **Complete**: Homework #11
	Unit 12: Data wrangling: string processing with stringr and parsing dates and times with lubridate	ULO 1: Use the stringrpackage to deal with strings that might be part of the original data or result from web-scraping the data. (SLO 1-3) ULO 2: Recognize the date as a data type. (SLO 1-3) ULO 3: Parse dates and times using lubridate. (SLO 1-3)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 24: String Processing b. Chapter 25: Parsing Dates and Times 2. **Complete**: Data Analysis Project #3 (ULO 1-3) 3. **Complete**: Homework #12 (ULO 1-3)
	Unit 13: Text mining	ULO 1: Process textual data using the tidytext package. (SLO 1-4) ULO 2: Summarize textual data using word clouds. (SLO 1-4) ULO 3: Extract information from textual data using topic modeling and sentiment analysis. (SLO 1-4)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 26: Text Mining 2. **Complete**: Homework #13 (ULO 1-3)
	Unit 14: Statistical learning: linear and logistic regressions	ULO 1: Explain the least squares estimation method. (SLO 5-6) ULO 2: Describe the purposes/uses of linear and logistic regression analysis. (SLO 5-6)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 17: Regression b. Chapter 18: Linear Models c. Chapter 27: Introduction to Machine Learning

		ULO 3: Develop and assess regression models for real datasets in R. (SLO 5-6) ULO 4: Use regression models for prediction. (SLO 5-6)	d. Chapter 31: Examples of Algorithms 2. **Complete**: Homework #14 (ULO 1-4)
	Unit 15: Statistical learning: classification and regression trees, and random forests	ULO 1: Describe pros and cons of tree-based methods for statistical learning. (SLO 5-6) ULO 2: Explain the process of building regression and classification trees. (SLO 5-6) ULO 3: Use random forests for prediction. (SLO 5-6)	1. **Read Textbook**: Irizarry, R. A. (2020). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R***.** Chapman and Hall/CRC a. Chapter 31: Examples of algorithms 2. **Complete**: Homework #15 (ULO 1-2) 3. **Complete**: Final Exam (ULO Unit 1-15) 4. **Complete**: Data Analysis Project Final (ULO 1-3)

** These descriptions and timelines are subject to change at the discretion of the instructor.*