DS 201
Introduction
to
Data Science

Instructor: Name

Contact: Email

Office Hours: Days/Times (Location)

Textbooks: *Data Science and Predictive Analytics* by Ivo D. Dinov (available electronically through Weinberg Memorial Library), *Introduction to Data Science* by Rafael A. Irizarry (available online `https://rafalab.github.io/dsbook/`), *R for Data Science* by Grolemund and Wickham (available online `https://r4ds.had.co.nz/`), and *An Introduction to Statistical Learning with Applicantions in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (available electronically through Weinberg Memorial Library)

Additional References: *Programming Skills for Data Science* by Michael Freeman and Joel Ross,*Probability and Statistics for Data Science* by Norman Matloff, and *Applied Predictive Modeling* by Max Kuhn and Kjell Johnson (available electronically through Weinberg Memorial Library)

Course Webpage:

Prerequisites:

Course Description: This course aims to provide students with an introduction to data science and the "data science" way of thinking. That is, we will explore principles for "drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference." Specifically, the course will introduce students to the basic data science workflow following current best practices, and address computational or algorithmic ways to think about data and learning from data. A heavy emphasis will be placed on data visualization, exploratory data analysis, and foundational modeling principles.

Course Goals: The goal of this course is for students

1. to gain confidence and foundational skills in implementing and applying the typical data science workflow: import, explore, format, transform, analyze, model, and draw conclusions from data.

2. to gain experience in effectively presenting and communicating results or findings based on an implementation or application of the typical data science workflow.

Student Learning Objectives and Assessment:

After taking this course, the student should be able to:	Methods of assessment
1. import, format, and transform common data-set types programmatically and build effective visualizations of data.	Programming competency assessments, homework, and project.
2. apply appropriate exploration and modeling techniques to learn from data.	Programming competency assessments, homework, and project.
3. present and communicate results obtained via data analysis in an effective manner.	Homework, and project.

Grade Policy: The course grade will be based on homework assignments (10%), programming competency assessments (10%), and a semester long project in the form of a data science case study (80%). The data science case study project will be broken up into a number of sub-assignments so that work on the project will span the entire semester, with various portions of the project due at different times throughout the semester.

Grading: Letter grades will be assigned based on the following scale:

Grade Range	Letter Grade
Below 60	F
$60 - 62$	D
$63 - 66$	$D+$
$67 - 69$	$C-$
$70 - 76$	C
$77 - 79$	$C+$
$80 - 82$	$B-$
$83 - 86$	B
$87 - 89$	$B+$
$90 - 92$	$A-$
$93 - 100$	A

Attendance: It is very important that you attend class regularly. In fact, much of the course content will be developed through outside assigned reading, whereas class time will be devoted largely to practice in the computer implementation of the various aspects of the data science workflow. Thus, it is the expectation of the instructor that students will participate in class activities. It is also extremely important for students to come prepared each day bringing their computers and having completed all assigned readings. Please be courteous and refrain from using your cell phone (talking or texting) or other electronic devices during class.

Homework: Students will be asked to complete a number of homework assignments throughout the semester. Homework will typically consist of conceptual questions requiring written responses or computational questions requiring students to write and/or implement code. The homework assignments are meant to assess understanding of and reinforce the assigned readings, and to provide practice in carrying out basic computations. Homework will make up 10% of the overall course grade and assignments will be posted on D2L.

Programming Competency Assessments: Throughout the semester students will be asked to demonstrate their competence in essential data science programming skills. At various times, students will be asked to meet individually with the instructor at which time the student will be asked to complete a pre-specified programming task. Students must successfully complete the assigned programming task, if a students fails to successfully complete the assigned task then they will be asked to return at a later time to make another attempt. A student successfully completing a task on the first try will be awarded ten points, one point will be deducted for each additional attempt required to successfully complete the task. The failure of a student to meet with the instructor in a timely fashion to complete a programming competency assessment will result in a grade of zero for the assignment. Programming competency assessments will make up 10% of the overall course grade.

Project: Students will develop a data science case study. The goal of the assignment is for students to implement the typical data science workflow on an approved data set(s) of their choosing. The final product will be a project notebook and a written report which includes a detailed explanation of the results obtained from an analysis of the data set(s) and a complete and professional description of the process used to obtain the presented results. All computer code must be thoroughly documented and included along with the submitted report. This project will make up 80% of the overall course grade. A detailed description of what constitutes a data science case study project and the requirements for completing the assignment will be posted on D2L. The data science case study project will be broken up into a number of sub-assignments. The various parts of the assignment together with their grade proportions and tentative due dates are:

1. selection and approval of a data set(s) - 5% (due $\approx$ September 13)
2. written data description - 5% (due $\approx$ September 20)
3. background, problem description, and objectives - 10% (due $\approx$ October 4)
4. programmatically loading, formatting, and transforming of the data set(s) - 12% (due $\approx$ October 11)
5. data visualization and summarization - 12% (due $\approx$ October 25)
6. modeling of the data - 12% (due $\approx$ November 15)
7. model evaluation - 12% (due $\approx$ November 22)
8. conclusions and report initial draft - 12% (due $\approx$ December 6)
9. report final version - 20% (due $\approx$ December 11)

Students will be allowed to revise earlier parts of the project assignment up until the final draft submission in order to receive (some but not all) points back.

Tentative Course Schedule:

1. Answer(s) to "What is data science?" ($\approx$ 1-2 classes)
2. Data, what it is, where to get it, different types, common data terminology and conventions ($\approx$ 2-3 classes)
3. Tools of the trade, i.e. essential computation, programming, development and version control ($\approx$ 5-6 classes)
4. Importing and exporting data ($\approx$ 2 classes)
5. Simple visualization ($\approx$ 3 classes)
6. Wrangling, i.e. formatting and working with the data ($\approx$ 3 classes)
7. Exploratory data analysis (EDA) and simple stats, *e.g.* essential background in correlation and regression ($\approx$ 4 classes)
8. Modeling ($\approx$ 12 classes)
9. On presentation of results and projects ($\approx$ 6 classes)

Important Dates:

Classes begin .....
Last day to add classes .....
Last day for 100% tuition refund .....
Last day to drop with no grade .....
Semester midpoint .....
Last day to withdraw with "W" .....