STAT 2600 – Introduction to Data Science

COURSE OBJECTIVES: This class will introduce the foundational knowledge, skills, and abilities of a modern data scientist. Specifically, students will learn to:

- locate, import, and simulate diverse data sets from multiple source types.
- aggregate, organize, summarize, and clean multiple data sets using wrangling techniques.
- construct, assess, and present insightful visualizations for a wide variety of data types.
- conduct exploratory analyses to discover and explain distributions and associations.
- compute and interpret inferential statistics via confidence intervals and hypothesis tests.
- build and assess the accuracy of predictive models for regression and classification.
- discuss and avoid ethical pitfalls in the collection, storage, and presentation of data.
- execute all the above learning objectives in the context of real-world problems.
- execute all the above learning objectives using the R coding language.

TEXTBOOK: *Data Science in Action*, 1st edition by Kristopher Pruitt. We will cover Chapters 1-5. Access to this electronic textbook will be provided to students free of charge.

SCHEDULE AND TOPICS COVERED

Lesson	Section	Topics
1	1.1	Data Science
2	1.2	Knowledge, Skills, and Abilities
3	1.3	The 5A Method
4	2.1	Data Structures
5	2.2.1	Importing Local Data
6	2.2.2	Importing Online Data
7	2.2.3	Simulating Data
8	2.3.1	Organizing Data
9	2.3.2	Summarizing Data
10	2.3.3	Aggregating Data
11	2.3.4	Cleaning Data
12	**Exam**	**Chapters 1-2**
13	3.1	Data Summary
14	3.2.1	Bar Charts
15	3.2.2	Histograms
16	3.2.3	Box Plots
17	3.2.4	Heat Maps
18	3.2.5	Contour Plots
19	3.3.1	Contingency Tables
20	3.3.2	Scatter Plots (Linear)
21	3.3.3	Scatter Plots (Nonlinear)
22	3.3.4	Scatter Plots (Logistic)
23	3.3.5	Line Graphs
24	4.1	Data Sampling
25	4.2.1	Sampling Distributions
26	4.2.2	Confidence Intervals (Proportion)
27	4.2.3	Confidence Intervals (Mean)
28	4.2.4	Confidence Intervals (Slope)
29	**Exam**	**Chapters 3-4.2**
30	4.3.1	Null Distributions
31	4.3.2	Hypothesis Tests (Proportion)
32	4.3.3	Hypothesis Tests (Mean)

33	4.3.4	Hypothesis Tests (Slope)
34	5.1	Data Modeling
35	5.2.1	Simple Linear Regression
36	5.2.2	Regression Accuracy
37	5.2.3	Multiple Linear Regression
38	5.2.4	Regression Trees
39	5.3.1	Simple Logistic Regression
40	5.3.2	Classification Accuracy
41	5.3.3	Multiple Logistic Regression
42	5.3.4	Classification Trees
43	**Exam**	**Chapters 4.3-5**

PREREQUISITES: None

EQUIVALENT COURSES: None

LEARNING OBJECTIVES BY SECTION

Lessons	Topics	Learning Objectives
1-3	Data Science	- Highlight the key historical trailblazers and milestones in the discipline of data science. - Describe the fundamental academic knowledge required of professional data scientists. - Detail the current technical and interpersonal skills needed to conduct data science. - Suggest how the abilities of data scientists can be applied in multiple, diverse domains. - Define the key characteristics and ethical considerations of good research questions. - Explain important attributes and collection methods for high-quality data and its sources. - Distinguish between goals and methods for exploratory, inferential, and predictive analyses. - Specify technical and non-technical considerations when advising stakeholders on results. - Define the key characteristics and ethical considerations for good research answers.
4-11	Data Acquisition	- Define the common terminology of data structures and apply it to real-world data. - Diagnose messy data structures and implement the required steps to make it tidy. - Identify appropriate variable types and sub-types for the columns of a data set. - Assign suitable primitive data types to variable values in a coding environment. - Determine appropriate non-primitive data types for groups of variable values. - Locate and import data into a coding environment from local and online sources. - Simulate data from the Binomial, Uniform, and Normal probability distributions. - Organize data by filtering, sorting, creating, and deleting rows and columns. - Summarize data via counts, proportions, sums, and averages of variable values. - Group data by factors and compute common summaries within the associated levels. - Join data frames by selecting the correct type of aggregation and primary key. - Compare and contrast observations between two data frames using join functions. - Find and resolve inconsistencies in naming variables and the levels of a factor. - Identify and resolve duplicated or missing observations and variable values.
12	Exam 1	
13-23	Exploratory Analyses	- Identify the correct variable types and associated visual cues in a data graphic. - Recognize the coordinate systems, scales, and units for variables in a data graphic. - Interpret the context in data graphics based on labels, captions, and annotations. - Identify and avoid (or resolve) common pitfalls in the visual presentation of data. - Summarize the distributions of variables using common measures of center and dispersion. - Interpret the distributions of categorical variables using bar charts and heat maps. - Interpret the distributions of numerical variables using histograms and contour plots. - Identify and resolve statistical outliers based on common statistics and box plots. - Summarize associations between factors with contingency tables and stacked bar charts. - Compute and interpret joint, marginal, and conditional proportions from a table. - Interpret the shape, direction, and strength of associations using scatter plots. - Apply linear transformations to nonlinear associations identified in a scatter plot. - Estimate and interpret the functional parameters for an identified association. - Characterize time-based associations between variable values using line graphs.

24-28	Inferential Analyses	- Summarize the common types of random sampling and their appropriate applications. - Define the common types of sampling bias and how each can be avoided or remedied. - Distinguish between and compute common population parameters and sample statistics. - Describe the key components of a confidence interval and their impact on its width. - Explain and execute bootstrap resampling with replacement based on a random sample. - Construct bootstrap sampling distributions and visualize confidence intervals or bounds. - Construct confidence intervals or bounds for a population proportion, mean, or slope. - Interpret intervals or bounds for a proportion, mean, or slope in real-world context.
29	Exam 2	
30-33	Inferential Analyses	- Describe the key steps of a hypothesis test and their impact on the conclusion. - Distinguish between one and two-sided hypotheses, and their impact on significance. - Construct a simulated null distribution and visualize the p-value and significance. - Complete hypothesis tests of a claim regarding a population proportion, mean, or slope. - Interpret hypothesis test results for a proportion, mean, or slope in real-world context.
34-42	Predictive Analyses	- Differentiate between supervised and unsupervised statistical learning based on objectives. - Identify the requirement for regression versus classification models based on the response. - Recognize parametric versus nonparametric modeling approaches based on the algorithm. - Explain the concept of over-fitting and how it is reduced by the validation set approach. - Split a random sample into a training set and a testing set using common ratios. - Estimate linear and logistic regression models with numerical predictors using training data. - Evaluate the accuracies of linear and logistic regression models using testing data. - Classify observations using predicted probabilities from a multiple logistic regression model. - Explain the structure and methods for developing regression and classification trees. - Build and assess decision trees to predict a response using recursive binary splitting.
43	Exam 3	