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, associations, and trends.
- 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, modeling, 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 for All*, 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

Day	Section	Topics
1	1.1	Knowledge, Skills, and Abilities
2	1.2	The 5A Method
3	1.3	Statistical Modeling
4	2.1	Structured Data
5	2.2	Importing Data
6	2.3	Simulating Data
7	2.4	Wrangling Data
8	2.5	Joining Data
9	2.6	Cleaning Data
10	2.7	Visualizing Data
11	Exam 1	Chapters 1 and 2
12	3.1	Categorical Variables
13	3.2	Contingency Tables
14	3.3	Text Mining
15	3.4	Numerical Variables
16	3.5	Outliers
17	3.6	Geospatial Analysis
18	3.7	Linear Association
19	3.8	Nonlinear Association
20	3.9	Time Series Analysis
21	3.10	Cluster Analysis
22	3.11	Principal Components
	Project 1	Chapter 3
23	4.1	Sampling and Bias
24	4.2	Confidence Intervals
25	4.3	Hypothesis Tests
26	4.4	Single Proportion
27	4.5	Two Proportions
28	4.6	Many Proportions
29	4.7	Single Mean
30	4.8	Two Means
31	4.9	Many Means

32	4.10	Single Slope
33	4.11	Technical Conditions
34	Exam 2	Chapter 4
35	5.1	Training and Testing
36	5.2	Bias-Variance Tradeoff
37	5.3	Cross-Validation
38	5.4	Multiple Linear Regression
39	5.5	Regression Accuracy
40	5.6	Decision Trees
41	5.7	Multiple Logistic Regression
42	5.8	Classification Accuracy
43	5.9	K-Nearest Neighbors
	Project 2	Chapter 5

PREREQUISITES: None

EQUIVALENT COURSES: None

LEARNING OBJECTIVES BY SECTION

Section	Topics	Learning Objectives
1.1	Knowledge, Skills, and Abilities	- 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.
1.2	The 5A Method	- 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.
1.3	Statistical Modeling	- 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.
2.1	Structured Data	- Define and apply the common terminology of data structures to a real-world data set. - Recognize messy data structures and reshape a real-world data set to make it tidy. - Assign primitive data types to the values of variables within a real-world data set. - Determine the appropriate type of non-primitive data based on a data set's dimensions.
2.2	Importing Data	- Locate, import, and structure data stored in a wide variety of built-in and local file types. - Locate, import, and structure data from a wide variety of online and remote sources. - Identify the appropriate variable types to assign to columns of a real-world data set.
2.3	Simulating Data	- Simulate discrete data from the binomial distribution given appropriate parameter values. - Simulate discrete data from the uniform distribution given appropriate parameter values. - Simulate continuous data from the normal distribution given appropriate parameter values.
2.4	Wrangling Data	- Organize a data frame by filtering, sorting, creating, and deleting its rows and columns. - Summarize a data frame via counts, proportions, sums, and averages of its variable values. - Group data by factor levels and compute summaries within the associated categories.
2.5	Joining Data	- Recognize the need for left, anti, or inner join types based on the purpose for aggregation. - Join multiple data frames by selecting the appropriate join type and primary keys. - Compare and contrast common observations between two tables using join functions.
2.6	Cleaning Data	- Find and resolve inconsistencies in naming variables and factor levels within a data frame. - Identify, investigate, and resolve duplicated observations and variables within a data frame. - Identify, investigate, and resolve missing or impossible variable values within a data frame.
2.7	Visualizing Data	- Identify the variable types present in a data graphic along with their associated visual cue. - Recognize the coordinate system, scales, and units of measurement within a data graphic. - Interpret the context of data graphics based on titles, labels, captions, and annotations.
3.1	Categorical Variables	- Summarize the distribution of a categorical variable using counts and proportions. - Visualize and interpret the distribution of a categorical variable using bar charts.

		- Identify and avoid common pitfalls in the presentation of categorical data in bar charts.
3.2	Contingency Tables	- Summarize the association between categorical variables using a contingency table. - Compute and interpret joint, marginal, and conditional proportions from a table. - Visualize and interpret associations between categorical variables using stacked bar charts.
3.3	Text Mining	- Define and correctly apply the key terminology and tools associated with text mining. - Summarize word and sentiment frequency within free text compilations using lexicons. - Visualize and interpret the evolution of sentiment within a text using line graphs.
3.4	Numerical Variables	- Summarize the distribution of a numerical variable based on its centrality and spread. - Visualize and interpret the distribution of a numerical variable using histograms. - Identify and avoid common pitfalls in the presentation of numerical data in histograms.
3.5	Outliers	- Summarize the distribution of a numerical variable using quartiles and interquartile range. - Visualize and compare the distributions of numerical variables using box plots. - Identify and resolve statistical outliers based on interquartile range and context.
3.6	Geospatial Analysis	- Visualize and interpret the geographic distribution of variables using choropleth maps. - Recognize and avoid issues related to color palette choice, specifically color blindness. - Create and apply custom functions to compute uncommon descriptive statistics.
3.7	Linear Association	- Characterize the linear association between two numerical variables using scatter plots. - Estimate and interpret parameters for the line of best fit using simple linear regression. - Explain the causes and implications for Simpson's Paradox using a colored scatter plot.
3.8	Nonlinear Association	- Characterize nonlinear associations between binary/numerical variables using scatter plots. - Estimate and interpret parameters for the curve of best fit using simple logistic regression. - Explain the purpose and benefits of the log-odds transformation of a logistic function.
3.9	Time Series Analysis	- Characterize the linear association between time-lagged variables using line graphs. - Estimate and interpret parameters for the trend of best fit using first order autoregression. - Identify and avoid common pitfalls in the presentation of time series data in line graphs.
3.10	Cluster Analysis	- Visualize and interpret clustering patterns among numerical variables using scatter plots. - Identify common subgroups among variables using K-means clustering techniques. - Explain the technical and practical implications of choosing a particular value for K.
3.11	Principal Components	- Recognize association between pairs of numerical variables using a scatter plot matrix. - Visualize and interpret principal components among variables using scatter plots. - Identify the principal components by computing the direction of maximum variance.
4.1	Sampling and Bias	- Summarize the common types of random sampling and their appropriate applications. - Define the common types of sampling bias and how each can be avoided. - Distinguish between and compute common population parameters and sample statistics.
4.2	Confidence Intervals	- 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 a bootstrap sampling distribution and use it to visualize a confidence interval.
4.3	Hypothesis Tests	- Describe the key steps of a hypothesis test and their impact on the conclusion. - Explain and execute randomization without replacement based on a random sample. - Construct a simulated null distribution and use it to visualize the p-value and significance.
4.4	Single Proportion	- Construct confidence intervals to estimate the true value of a single proportion. - Complete a hypothesis test of a claim regarding the true value of a single proportion. - Interpret inferences on a single proportion in the context of a real-world problem.
4.5	Two Proportions	- Construct confidence intervals to estimate the difference between two proportions. - Complete a hypothesis test of a claim regarding the difference between two proportions. - Interpret inferences on differences in proportions in the context of a real-world problem.
4.6	Many Proportions	- Complete a goodness-of-fit hypothesis test for the distribution of many proportions. - Interpret tests of the distribution of proportions in the context of a real-world problem.
4.7	Single Mean	- Construct confidence intervals to estimate the true value of a single mean. - Complete a hypothesis test of a claim regarding the true value of a single mean. - Interpret inferences on a single mean in the context of a real-world problem.
4.8	Two Means	- Construct confidence intervals to estimate the difference between two means. - Complete a hypothesis test of a claim regarding the difference between two means. - Interpret inferences on differences in means in the context of a real-world problem.
4.9	Many Means	- Complete an analysis of variance hypothesis test of the difference between many means. - Interpret analysis of variance tests for many means in the context of a real-world problem.
4.10	Single Slope	- Construct confidence intervals to estimate the true value of a single slope parameter. - Complete a hypothesis test of a claim regarding the true value of a single slope parameter. - Interpret inferences on a single slope parameter in the context of a real-world problem.
4.11	Technical Conditions	- Diagnose issues with the technical conditions of linear regression via diagnostic plots. - Remedy issues with the technical conditions of linear regression via transformation.

5.1	Training and Testing	- Explain the concept of overfitting and how it is remedied by the validation set approach. - Visualize the impact of model flexibility on training and testing errors using a line graph. - Split a random sample into a training set and a testing set using common ratios.
5.2	Bias-Variance Tradeoff	- Describe the bias-variance trade-off inherent in developing statistical learning models. - Visualize the impact of model flexibility on bias and variance using a line graph. - Distinguish modeling approaches with high bias from those with high variance.
5.3	Cross-Validation	- Explain the structure and benefits of cross-validation given the bias-variance trade-off. - Execute LOO cross-validation to estimate the accuracy of a simple linear regression. - Execute k-fold cross-validation to estimate the accuracy of a simple linear regression.
5.4	Multiple Linear Regression	- Construct a multiple linear regression model with only continuous numerical predictors. - Assess the quality of fit for multiple linear regression models using residual standard error. - Compute confidence intervals for the numerical response using bootstrap resampling.
5.5	Regression Accuracy	- Evaluate the prediction accuracy of a multiple linear regression model using test data. - Choose between two multiple linear regression models based on cross-validated error.
5.6	Decision Trees	- Explain the structure and methods for developing tree-based regression models. - Build and assess regression trees to predict a response using recursive binary splitting. - Articulate the pros and cons of non-parametric methods versus linear regression.
5.7	Multiple Logistic Regression	- Construct a multiple logistic regression model with only continuous numerical predictors. - Assess the quality of fit for multiple logistic regression models using residual deviance. - Compute confidence intervals for the probability of success using bootstrap resampling.
5.8	Classification Accuracy	- Classify observations using predicted probabilities from a multiple logistic regression. - Evaluate the classification accuracy of a multiple logistic regression model using test data. - Choose between two multiple logistic regression models based on cross-validated error.
5.9	K-Nearest Neighbors	- Explain the structure and methods for creating K-nearest neighbors classification models. - Build and assess a K-nearest neighbors model to classify observations in a test set. - Articulate the pros and cons of non-parametric methods versus logistic regression.