General Assembly
Course Curriculum

DATA SCIENCE

GA.CHI/DS

Data Science
Table of Contents

OVERVIEW

THE FRAMEWORK

Ever wonder how the Netflix recommendation engine works or how Amazon.com determines what items "you may also like?" All of these things are driven by training a computer how to learn using the large datasets.

The data science course is a practical introduction to the interdisciplinary field of data science and machine learning which is at the intersection of computer science, statistics, and business. You will learn to use Python to help you acquire, parse and model your data. A significant portion of the course will be a hands-on approach to the fundamental modeling techniques and machine learning algorithms that enable you to build robust predictive models of real-world data and test their validity. You will also practice communicating your results and insights. By the end of the course, students will be able to:

- Perform exploratory data analysis with python.
- Build and refine machine learning models to predict patterns from data sets.
- Communicate data driven insights to a technical and non-technical audience alike.

STUDENTS

DATA ANALYSTS OR BUSINESS INTELLIGENCE ANALYSTS

This course provides data professional with the skills required to solve problems using computation that involve large data sets such as predicting user behavior on their website, making decisions, or the best way to classify content. Individuals learn how to build the code necessary to be able to make predictions and create models.

PROJECTS

UNIT PROJECTS

Globally, we have 4 Unit Projects in Data Science, each building on top of skills learned previously to scaffold students' learning over the entire course. Our projects include objectives, requirements, starter-code, rubric, and suggested resources - all of which tie into the overall competencies for each unit.

FINAL PROJECT

For the Data Science final project, you will address a data-related problem in your professional field or in a field you interested in. You will acquire a real-world data set, form a hypothesis about it, clean, parse, and apply modeling techniques and data analysis principles to ultimately create a predictive model. Students present their results and each write a report that includes the following:

- Clearly articulated problem statement
- Summary of data acquisition, cleaning, and parsing stage
- Clear presentation of your predictive model and the processes you took to create it
- Presentation style appropriate to the audience

Your instructor will help you scope out your project so that you choose something that is feasible to accomplish given the skills you acquire in the course.

PROJECT TIMELINE

UNIT	Project	Assigned	Deadline
Unit 1	Project 1	Lesson 2	Lesson 4
Unit 1	Project 2	Lesson 4	Lesson 6
Unit 2	Final Project, pt 1	Lesson 1	Lesson 8
Unit 2	Project 3	Lesson 5	Lesson 10
Unit 2	Project 4	Lesson 9	Lesson 12
Unit 3	Final Project, pt 2	Lesson 8	Lesson 14
Unit 3	Final Project, pt 3	Lesson 14	Lesson 16
Unit 3	Final Project, pt 4	Lesson 16	Lesson 18
Unit 3	Final Project, pt 5	Lesson 18	Lesson 20

UNITS

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	What is Data Science	Lesson 1
	Research Design and Pandas	Lesson 2
	Online Data Acquisition	Lesson 3
	Statistics Fundamentals I	Lesson 4
	Statistics Fundamentals II	Lesson 5
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

1 RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

1 WHAT IS DATA SCIENCE

- Describe course syllabus and setup development environment
- Answer the questions: "What is Data Science? What roles exist in Data Science?"
- Define the workflow, tools and approaches data scientists use to analyze data

2 RESEARCH DESIGN AND PANDAS

- Define a problem and identify appropriate data sets using the data science workflow
- Walkthrough the data science workflow using a case study in the Pandas library
- Import, format and clean data using the Pandas Library

3 DATA ACQUISITION

- Use Python to download datasets from the internet
- Connect to APIs and send data requests
- Parse structured and unstructured data available online

4 STATISTICAL FUNDAMENTALS I

- Use NumPy and Pandas libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile, range, variance, standard deviation and correlation
- Create data visualization - scatter plots, scatter matrix, line graph, box blots, and histograms- to discern characteristics and trends in a dataset
- Identify a normal distribution within a dataset using summary statistics and visualization

5 STATISTICAL FUNDAMENTALS II

- Explain the difference between causation vs. correlation
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (p-values, confidence intervals)

2 FOUNDATION OF DATA MODELING

6 LINEAR REGRESSION

- Define data modeling and linear regression
- Differentiate between categorical and continuous variables
- Build a linear regression model using a dataset that meets the linearity assumption using the scikit-learn library

7 EVALUATING MODEL FIT

- Define regularization, bias, and errors metrics
- Evaluate model fit by using loss functions including mean absolute error, mean squared error, root mean squared error
- Select regression methods based on fit and complexity

8 INTRODUCTION TO CLASSIFICATION

- Define a classification model
- Build a K-Nearest Neighbors using the scikit-learn library
- Evaluate and tune model by using metrics such as classification accuracy/error

9 LOGISTIC REGRESSION I

- Build a Logistic regression classification model using the scikit-learn library
- Describe the sigmoid function, odds, and odds ratios and how they relate to logistic regression
- Evaluate a model using metrics such as classification accuracy/error, confusion matrix, ROC / AOC curves, and loss functions

10 COMMUNICATING LOGISTIC REGRESSION RESULTS

- Explain the tradeoff between the precision and recall of a model and articulate the cost of false positives vs. false negatives.
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders
- Describe the difference between visualization for presentations vs. exploratory data analysis

11 FLEXIBLE CLASS SESSION

- Focus on a topic selected by the instructor/class in order to provide deeper insight into data modeling

3 DATA SCIENCE IN THE REAL WORLD

12 DECISION TREES AND RANDOM FOREST

- Describe the difference between classification and regression trees and how to interpret these models
- Explain and communicate the tradeoffs of decision trees vs regression models
- Build decision trees and random forests

13 NATURAL LANGUAGE PROCESSING

- Demonstrate how to tokenize natural language text
- Categorize and tag unstructured text data
- Explain how to build a text classification model using NLTK

14 DIMENSIONALITY REDUCTION

- Explain how to perform a dimensional reduction Demonstrate how to refine data using Latent Dirichlet Allocation (LDA)
- Extract information from a sample text dataset

15 TIME SERIES DATA

- Explain why time series data is different than other data and how to account for it
- Create rolling means and plot time series data
- Perform autocorrelation on time series data

16 CREATE MODELS WITH TIME SERIES DATA

- Decompose time series data into trend and residual components
- Validate and cross-validate data from different data sets
- Use the ARIMA model to forecast and detect trends

17 DATABASE TECHNOLOGIES

- Describe the use cases for different types of databases
- Explain differences between relational databases and document-based databases
- Write simple select queries to pull data from a database and use within Pandas

18 WHAT'S NEXT?

- Identify next steps in data science learning

19/20 FINAL PROJECT PRESENTATIONS

- Final project presentation and discussion

FAQS

WHY IS THIS COURSE RELEVANT TODAY?

Given the large amount of data available, businesses could be making more data driven decisions if this vast amount of data was more deeply analyzed through the use of data science. The data science course provides the tools, methods, and practical experience to enable you to make accurate predictions about data, which ultimately leads to better decision-making in business, and the use of smarter technology.

WHAT PRACTICAL SKILL SETS CAN I EXPECT TO HAVE UPON COMPLETION OF THE COURSE?

This course provides you with technical skills in machine learning, algorithms, and data modeling which allow you to make accurate predictions about your data. You'll create your models using Python. Furthermore, you will learn how to programmatically parse and clean your data.

WHO WILL I BE SITTING NEXT TO IN THIS COURSE?

Individuals who have a strong interest in manipulating large data sets, finding patterns in data, and making predictions. Analysts and Business Intelligence Analysts who want to level up their skill set with data modeling. Individuals with a good grasp of data, a solid knowledge of statistics and probability. Pre-work:

- CodeAcademy: Learn Python

FAQS

WHAT CAN I EXPECT BY THE END OF THE COURSE?

By the end of the course, you can expect to be able to acquire, parse, clean, and apply various modeling techniques to your data to make predictions. You should also be able to communicate your findings to both a non-technical and technical audience in both written and verbal formats.

WILL THERE BE ANY PRE-WORK?

Yes. You will be required to complete approximately 10 - 15 hours of pre-work.

SHOULD I COME EQUIPPED WITH ANYTHING?

Yes. Please come prepared with a laptop (Mac OSX is preferred but not required).

CONTACT

Instructor: Ivan Hernandez, Ph.D

Email: Ivan@ivanhernandez.com

Slack: ivanhrndz

Office hours: Flexible (e-mail or ask after class to set up a time at your convenience)