General Assembly
Course Curriculum

DATA SCIENCE

GA.CO/DS

Data Science
Table of Contents

3 Overview

4 Students

5 Curriculum
Projects & Units

10 Frequently
Asked Questions

12 Contact
Information

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

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 instructional team 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.

UNITS

UNIT 1: FOUNDATIONS OF DATA MODELING	Introduction to Data Science	Lesson 1
	Elements of Data Science	Lesson 2
	Data Visualisation	Lesson 3
	Linear Regression	Lesson 4
	Logistic Regression	Lesson 5
	Model Evaluation	Lesson 6
	Regularisation	Lesson 7
	Clustering	Lesson 8
UNIT 2: DATA SCIENCE IN THE REAL WORLD	Recommendations	Lesson 9
	SQL + Productivity	Lesson 10
	Decision Trees	Lesson 11
	Ensembles	Lesson 12
	Natural Language Programming	Lesson 13
	Cloud Computing	Lesson 14
	Time Series	Lesson 15
	Soft Skills	Lesson 16
	Network Analysis	Lesson 17
	Neural Networks	Lesson 18
	Final Projects Presentations	Lesson 19
	Final Projects Presentations	Lesson 20

UNITS

UNIT 1: FOUNDATIONS OF DATA MODELING

1 INTRODUCTION TO 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 ELEMENTS OF DATA SCIENCE

- 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 VISUALISATION

- Create data visualisation - 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 visualisation
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders
- Describe the difference between visualisation for presentations vs. exploratory data analysis

4 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

5 LOGISTIC REGRESSION

- 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

6 MODEL EVALUATION

- Evaluate a model using metrics such as classification accuracy/error, confusion matrix, ROC / AOC curves, and loss functions
- 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
- Explain the tradeoff between the precision and recall of a model and articulate the cost of false positives vs. false negatives

7 REGULARISATION

- Define regularisation, bias, and errors metrics
- Explain why regularisation is used
- Use regularisation to reduce model complexity

8 CLUSTERING

- 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

UNIT 2: DATA SCIENCE IN THE REAL WORLD

9 RECOMMENDATIONS

- Explain recommendation engines and their functionality
- Explain the difference between content-based filtering and collaborative filtering
- Build a recommendation engine

10 SQL + PRODUCTIVITY

- 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

11 DECISION TREES

- 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

12 ENSEMBLES

- Explain ensemble methods and why they are used
- Explain bagging and boosting ensemble methods
- Build a random forest model

13 NATURAL LANGUAGE PROGRAMMING

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

14 CLOUD COMPUTING

- Explain what cloud computing is and how it is used
- Explore cloud computing options, including Spark

15 TIME SERIES

- 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
- 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

16 SOFT SKILLS

- Discuss the importance of effective communication in data science
- Outline the requirements of a successful presentation
- Appreciate the complexity and importance of organisational politics
- Describe techniques for effective consulting
- Explain the role of governance in data and analytics
- Approach the job market with a more focused strategy

17 NETWORK ANALYSIS

- Explain what a network is
- Discuss basic graph theory
- Describe properties and metrics of graphs
- Perform network analysis using Python

18 NEURAL NETWORKS

- Explore neural networks and what they involve

19/20 FINAL PROJECTS PRESENTATIONS

- Identify next steps in data science learning
- 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

COURSE APPLICATION

NEW YORK
nyc_admissions@ga.co
1 (877) 348-5665

BOSTON
boston_admissions@ga.co
1 (877) 348-5665

WASHINGTON, DC
dc_admissions@ga.co
1 (877) 348-5665

CHICAGO
chicago_admissions@ga.co
1 (877) 348-5665

ATLANTA
atlanta_admissions@ga.co
1 (877) 348-5665

SEATTLE
seattle_admissions@ga.co
1 (877) 348-5665

AUSTIN
austin_admissions@ga.co
1 (877) 348-5665

SAN FRANCISCO
sf_admissions@ga.co
1 (877) 348-5665

LOS ANGELES
la_admissions@ga.co
1 (877) 348-5665

LONDON
london_admissions@ga.co
+44 (0) 7703 002 046

SYDNEY
sydney_admissions@ga.co
+61 2 0420 438 649

HONG KONG
hk_admissions@ga.co
+852 3952 7209

MELBOURNE
melbourne_admissions@ga.co
+61 1300 779 650

SINGAPORE
singapore_admissions@ga.co