American Public University System
American Military University | American Public University

# DATS211

## Course Summary

**Course :** DATS211 **Title :** Introduction to Data Science
**Length of Course :** 8 Weeks
**Prerequisites:** **MATH302 or statistics equivalent, MATH220 or linear algebra equivalent, and DATS201 Analytical Methods I** **Credit Hours :** 3

## Description

This course provides an overview of data science including a foundation in research methodology. Data science is a data-driven process that provides descriptive, predictive and prescriptive insight. Whether reporting on historical information or making predictions about future events, the goal of data science is to add value through analysis that informs. To meet this goal this course introduces a range of tools and methods including supervised and unsupervised techniques. These include techniques such as classification, rule-based association techniques, support vector machines, K-nearest neighbor, regression, and clustering techniques such as K-Means.

## Course Scope:

This course is intended for undergraduate students studying data science. It provides students with an introduction to descriptive and predictive analyses. For example, descriptive analyses include discovering information about the data available to solve a problem or address a challenge, e.g. statistical information such as means, standard deviations, etc. It also includes initial visualization of data such as histograms, scatter plots and boxplots. Predictive analyses can include inference methods, hypothesis testing, building linear models, and more. Students will complete this course with a good understanding of the basic data analyses conducted in data science.

## Objectives

At the conclusion of this course, students will be able to:

L01: Demonstrate an understanding of the underlying methods and techniques used in data science.

L02: Given a specified problem be able to:

- Describe the methods and techniques that are appropriate for solving the problem
- Complete a data analysis that provides a solution for the problem
- Evaluate that analysis to determine its accuracy

- Describe methods to increase the accuracy of the analysis
- Draft a report including the results of the analysis, the evaluation of the analysis and recommendations such as how to increase the accuracy of the analysis.

**Textbook (Required)**
Free open source textbooks will be used throughout this course. This includes videos, course notes and a variety of articles. http://faculty.marshall.usc.edu/gareth-james/ISL/ and https://rafalab.github.io/dsbook/. Extra materials will be made available to students for contents outside the prescribed textbook.

**Software (Required)**
Required software includes the R programming language at: https://www.r-project.org/ and RStudio at: https://rstudio.com/, as well as RTools at: https://cran.r-project.org/bin/windows/Rtools/. Note that if you are using R 4.0.0 you will need to download and install RTools40 at: https://cran.r-project.org/bin/windows/Rtools/. You will need RTools or RTools40 to install the suite of "tidyverse" packages.

**Readings:**
**Machine Learning with R**, Brett Lantz, available online at the APUS/Trefry Library

and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select mac version)

**Other Resources (Not Required):**

**kaggle** Introduction to machine learning in R (tutorial) at: https://www.kaggle.com/camnugent/introduction-to-machine-learning-in-r-tutorial

**mlr.org** (Machine Learning in R) at: https://mlr.mlr-org.com/

**Engineering Statistics** The National Institute of Standards (NIST) available at: https://www.itl.nist.gov/div898/handbook/index.htm or,

**NIST/SEMATECH e-Handbook of Statistical Methods**, http://www.itl.nist.gov/div898/handbook/, July 29, 2020.

https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/index.htm#

https://www.oercommons.org/courses/prediction-machine-learning-and-statistics-spring-2012

**Software (Optional)**
1. Python version 3 on mac and windows available at: https://www.python.org/downloads/
2. Other software usage will require approval from the instructor.

# Outline

**Week 1: Data Science, Supervised and Unsupervised Methods, Working with data & Software Installations**

**Learning Outcomes:**

At completion of the modules this week students will be able to:

1. Install required software.
2. Distinguish between supervised and unsupervised methods based on data
3. Perform basic to intermediate programming in R.

**Module 1 – Unit 2.1-3.2:** Review of Data Science and methods

**Module 2 – Chapter 3 of** Wickham **R for Data Science:** Data Visualization

Extra hands-on programming experience in R:

If needed, review **Hands-On Programming with R** at: https://rstudio-education.github.io/hopr/,

Irizarry Intro to Data Science chapters as required, and/or Wickham **R for Data Science** Chapters 1 and 2

**Homework I: Use the data attached to answer the following questions:**

Reference Week 1 HW folder

**Week 2: Exploratory Data Analyses**

**Learning Outcomes:**

At completion of the modules this week students will be able to:

1. Understand the relevance of Exploratory Data Analyses (EDA) in data science.
2. Understand the role of descriptive data science.
3. Understand the role of predictive analyses in data science
4. Understand the differences between EDA, descriptive and predictive analytics

**Module 3**: Review Kaggle Introduction to machine learning in R (tutorial)

**Module 4**: "Exploratory Data Analysis" in the Engineering Statistics Handbook at: https://www.itl.nist.gov/div898/handbook/eda/eda.htm

Lab #1: Report on the differences between Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA)

**Week 3: Regression**

At completion of the modules this week students will be able to:

1. Prepare data for analyses.
2. Build regression models and interpret the outcome
3. Evaluate the model's performance
4. Incorporate one and two way interaction effect in the model

**Module 5 –** Irizarry **Introduction to Data Science** Chapter 17

**Module 6** - Follow the IDRE seminar: Introduction to Regression in R at: https://stats.idre.ucla.edu/r/seminars/introduction-to-regression-in-r/# (Institute for Digital Research & Education. UCLA: Statistical Consulting Group, 2020)

Reference Week 3 HW folder

**Week 4: Classification and Trees**

At completion of the modules this week students will be able to:

1. Train a classification model, validate and evaluate it.
2. Train, validate and evaluate decision tree models.

**Module 7-** Classification using nearest neighbors**.**

Resources http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/06_LazyLearning_kNN.html

**Module 8** – Classification using Naïve Bayes Classifications

Resources http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/07_NaiveBayesianClass.html

**Module 9-** Classification using decision tree models

Resource: http://www.socr.umich.edu/people/dinov/2017/Spring/DSPA_HS650/notes/08_DecisionTreeClass.html

Reference Week 4 HW folder

**Week 5: Clustering**

At completion of the modules this week students will be able to:

1. Gain further understanding of clustering methods.
2. Apply and evaluate KNN on real world data.
3. Apply and evaluate K-Means as a clustering model on real world data.
4. Identify the differences between the two methods at the theoretical and application levels.

**Module 10**: Review the information at the SOCRAT webpage http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/06_LazyLearning_kNN.html on clustering and k-Nearest Neighbors including the case study on the "Boys Town Study of Youth Development".

**Module 11**: Review the information at the SOCRAT webpage http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/12_kMeans_Clustering.html#4_case_study_1:_divorce_and_consequences_on_young_adults on k-Means clustering including the case study on "Divorce and Consequences on Young Adults".

Reference week 5 HW folder

**Week 6: Build Your Dashboard**

At completion of the modules this week students will be able to:

1. Develop a proposal about their dashboard projects.
2. Review documentation for the course projects and gather resources for the project.
3. Understand boosting and how it differs from bagging.
4. Begin the first steps in developing a dashboard for their chosen company.

**Module 12:** Review the material at: https://www.r-bloggers.com/creating-a-business-dashboard-in-r/

**Module 13**: Review the material from RStudio at: https://shiny.rstudio.com/articles/dashboards.html including the videos (Dynamic Dashboards, Building Dashboards, and Dashboards made easy)

https://github.com/rstudio-education/shiny.rstudio.com-tutorial/blob/master/how-to-start-shiny-part-1.pdf

**Module 14**: Additional material and an alternative approach is at: https://rmarkdown.rstudio.com/flexdashboard/

Reference app.R file in HW folder to use as a template if you wish.

**Week 7: Analytics**

At completion of the modules this week students will be able to:

1. Understand and apply model evaluation technique to solve real world problems.
2. Understand model performance and how to assess it.

**Module 15**: Review Chapter 13 from the SOCRAT webpage
http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/13_ModelEvaluation.html

**Module 16**: "Various ways to evaluate a machine learning model's performance" by Kartik Nighania at:
https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Module 17: "Why Question Machine Learning Evaluation Methods?" by Nathalie Japkowicz at:
https://www.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-003.pdf (Note that aaai.org is the Association for the Advancement of AI.)

Reference Week 7 HW folder

**Week 8: Final Project Presentation**

At completion of the modules this week students will be able to:

1. Present their projects thoroughly while taking in feedback to improve future work
2. Use and be comfortable with presenting data science concepts and their application in the project.
3. Draw effectively from concepts covered in the class.

# Evaluation

A variety of weekly exercises (Knowledge Checks) will be used to reinforce the material covered in this course. This course will include a final exam, weekly homework, quizzes, discussions and labs (that are not required but highly recommended). The grade distribution of the course is shown below.

- Homework 25%
- Laboratories 15%
- Discussions 10%
- Knowledge Checks 20%
- Final Presentation 30%

# Materials

**Textbook (Required)**

Free open source textbook will be used throughout this course. This includes videos, course notes and a variety of articles. http://faculty.marshall.usc.edu/gareth-james/ISL/ . Extra materials will be made available to students for contents outside the prescribed textbook.

**Software (Required)**

1. Install R and RStudio on windows from: https://cran.r-project.org/bin/windows/base/ and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select windows version)
2. Install R and Rstudio on mac from: https://cran.r-project.org/bin/macosx/ and https://rstudio.com/products/rstudio/download/ (scroll down to All installers and select mac version)

**Software (Optional)**

3. Python version 3 on mac and windows available at: https://www.python.org/downloads/
4. Other software usage will require approval from the instructor.

# Late Assignments

1. Students are expected to submit classroom assignments by the posted due date and to complete the course according to the published class schedule. The due date for each assignment is listed under each Assignment.
2. Generally speaking, late work may result in a deduction up to 15% of the grade for each day late, not to exceed 5 days.
3. As a working adult I know your time is limited and often out of your control. Faculty may be more flexible if they know ahead of time of any potential late assignments.

# Policies

Please see the Student Handbook to reference all University policies. Quick links to frequently asked question about policies are listed below.

Drop/Withdrawal Policy

Plagiarism Policy

Extension Process and Policy

Disability Accommodations

**Writing Expectations**

All written submissions should be submitted in a font and page set-up that is readable and neat. It is recommended that students try to adhere to a consistent format, such as that described below.

- Typewritten in double-spaced format with a readable style and font and submitted inside the electronic classroom (unless classroom access is not possible and other arrangements have been approved by the professor).
- 11 or 12-point font in a style such as Arial, Helvetica or Times New Roman.

**Citation and Reference Style**

Assignments completed in a narrative essay or composition format must follow a widely accepted citation style, such as APA, Turabian or MLA. Please refer to the APUS Online Library for further examples, or contact the instructor with questions.

**Late Assignments**

Students are expected to submit classroom assignments by the posted due date and to complete the course according to the published class schedule. As adults, students, and working professionals, I understand you must manage competing demands on your time. Should you need additional time to complete an assignment, please contact me **before the due date** so we can discuss the situation and determine an acceptable resolution. Routine submission of late assignments is unacceptable and may result in points deducted from your final course grade.

**Netiquette**

Online universities promote the advancement of knowledge through positive and constructive debate – both inside and outside the classroom. Forums on the Internet, however, can occasionally degenerate into needless insults and "flaming." Such activity and the loss of good manners are not acceptable in a university setting – basic academic rules of good behavior and proper "Netiquette" must persist. Remember that you are in a place for the rewards and excitement of learning which does not include descent to personal attacks or student attempts to stifle the Forum of others.

- **Technology Limitations:** While you should feel free to explore the full-range of creative composition in your formal papers, keep e-mail layouts simple. The Sakai classroom may not fully support MIME or HTML encoded messages, which means that bold face, italics, underlining, and a variety of color-coding or other visual effects will not translate in your e-mail messages.

- **Humor Note:** Despite the best of intentions, jokes and especially satire can easily get lost or taken seriously. If you feel the need for humor, you may wish to add “emoticons” to help alert your readers: ;-), : )

**Disclaimer Statement**

Course content may vary from the outline to meet the needs of this particular group.

# Online library

The Online Library is available to enrolled students and faculty from inside the electronic campus. This is your starting point for access to online books, subscription periodicals, and Web resources that are designed to support your classes and generally not available through search engines on the open Web. In addition, the Online Library provides access to special learning resources, which the University has contracted to assist with your studies. Questions can be directed to **librarian@apus.edu**.

- **Charles Town Library and Inter Library Loan:** The University maintains a special library with a limited number of supporting volumes, collection of our professors’ publication, and services to search and borrow research books and articles from other libraries.
- **Electronic Books:** You can use the online library to uncover and download over 50,000 titles, which have been scanned and made available in electronic format.
- **Electronic Journals:** The University provides access to over 12,000 journals, which are available in electronic form and only through limited subscription services.
- **Tutor.com**: AMU and APU Civilian & Coast Guard students are eligible for 10 free hours of tutoring provided by APUS. Tutor.com connects you with a professional tutor online 24/7 to provide help with assignments, studying, test prep, resume writing, and more. Tutor.com is tutoring the way it was meant to be. You get expert tutoring whenever you need help, and you work one-to-one with your tutor in your online classroom on your specific problem until it is done.

**Library Guide (http://apus.campusguides.com/SCIN134)**

The AMU/APU Library Guides provide access to collections of trusted sites on the Open Web and licensed resources on the Deep Web. This course guide provides links to a number of sources relevant to this course, including journals, books, and web sites. Also, you can directly contact the librarian assigned to this course for assistance in locating information.