ACM Data Science Task Force Course Example

Data Science in R
Northwestern CT Community College
Winsted, CT
Crystal Wiggins
cwiggins@nwcc.edu

Link to complete course materials:.
https://drive.google.com/drive/folders/1_B55H9rQkPbCDJA-NpwjYI893ktJcF0h?usp=sharing
.

Knowledge Areas that contain competencies (knowledge, skills, and dispositions) covered in the course

Knowledge Area	**Total Number of Contact Hours (credits)**
Applied Programming (PRG)	*2.75*
Statistical Programming (SP)	0.25

Where does the course fit in your undergraduate Data Science curriculum?
Data Science in R is a required course in the associate's degree and the certificate. This course is taken in the first year, depending on where they start in math, as Principles of Statistics is a prerequisite and/or corequisite.

Is this course from or used in other curricula/majors?
This course was designed for Data Science majors and as a sufficient programming course that can be used for Computer Science majors.

What is covered in the course?
Introduction to the field of data science and the programming language of R. Explores the data science lifecycle, including question formulation, data collection and cleaning, exploratory data analysis and visualization, statistical inference and prediction, and decision-making. Focuses on quantitative critical thinking and key principles and techniques needed to carry out this cycle. No prior programming experience required.

[For topics, please see Course Overview in the google drive folder. Link is above.]

What is the format of the course?
This course can be taught in all formats; online, hybrid, face-to-face. It is taught through project-based learning techniques. Students are taught how to find the answers they need as they

encounter the problem that needs solving. A main focus is teaching them that the fear of the unknown is a normal feeling and to embrace that fear. Students use a programming system (i.e. DataQuest) to learn the language and apply what they learned through projects. This course is 3 contact hours.

How are students assessed?
Every few weeks, there is a programming project assigned and a week is given to complete as much as they can. The goal of the projects is to build their confidence in programming and problem solving. There are no exams, everything is applied and applicable to the field of work.

Course tools and materials

- DataQuest – an online system that offers several programming tracks. They have a variety of packaging licenses. I worked with their administration for my students to receive full access to the site for the semester for free!
- R language is used, however python can be used as well.
- Medium-large data sets are used.
- R Studio (cloud or downloadable) - free

Why do you teach the course this way?

Goals:
Students will develop a strong foundation in the programming language of R, including the various concepts, methodologies, and competencies that a data scientist must possess in order to be successful. These include the data science lifecycle, decision-making, and quantitative critical-thinking.

Outcomes:
Upon successful completion of this course, each student will be able to:

1. Explain the field of data science.
2. Apply techniques to import, clean, and transform data.
3. Practice exploratory analysis and visualization of data techniques.
4. Analyze and interpret data to tell a story.
5. Utilize the programming language R to manipulate data.

This course was created and piloted in Fall 2019. Students found the course to be challenging but at an appropriate level. This course was designed using project-based learning. The lectures are focused on career-based information (i.e. LinkedIn, resume, interviews for a data science position, learning the field of data science). Lectures also introduce the use of R Studio and the different formats available. Projects are completed in R Studio. All of the programming is learned through the DataQuest platform and evaluated through the projects.

Body of Knowledge coverage
[For each Knowledge Area, list the sub-domains covered in whole or in part in the course. If in part, please indicate which knowledge/skills/dispositions are covered. This section will likely be the most time-consuming to complete, but is the most valuable for educators planning to adopt

the computing-specific recommendations for a Data Science program at the undergraduate level.]

KA	**Sub-domain**	**Competencies Covered**	**Hours (in a semester)**
PDA	*Data Analysis*	● Arithmetic Expressions and Variables ● Logical Expressions ● Data Manipulation (Basics)	2
PDA	Data Structures	● Vectors ● Matrices ● Lists ● Dataframes	3
PDA	Control Flow, Iteration, and Functions	● Control Flow ● Iterations ● Functions	2
PDA	Data Processing	● String Manipulation ● Data and time Manipulation ● The Map Function	2
AP	Data Visualization	● Creating Line Graphs ● Creating Multiple Line Graphs ● Bar Charts, Histograms, and Box Plots ● Scatter Plots for Exploratory Analysis	3
DG	Data Cleaning	● Data Cleaning (intro) ● Staring Manipulation and Relational Data ● Correlations and Reshaping Data ● Dealing with Missing Data ● Regular Expressions (basics) ● Advanced Regular Expressions ● Map and Anonymous Functions ● Working with Missing Data	9
PDA	SQL	● Intro to SQL ● Summary Statistics ● Group Summary Statistics ● Subqueries ● Joining Data ● Building and Organizing Complex Queries ● Querying SQLite ● Table Relations and Normalization	9

SP	Statistics Fundamentals in R	● Simple Random Sample ● Stratified Sampling and Cluster Sampling ● Variables in Statistics ● Frequency Distributions ● Visualizing Frequency Distributions ● Comparing Frequency Distributions	3
SP	Statistics Intermediate in R	● The Mean ● The Weighted Mean and the Median ● The Mode ● Measures of Variability ● Z-scores	3