Introductory Data Science: A Blueprint to Navigate Curricular, Pedagogical, and Computational Challenges

Elijah Meyer
Department of Statistics, Duke University
and
Mine Çetinkaya-Rundel
Department of Statistics, Duke University

January 17, 2023

Abstract

The text of your abstract. 200 or fewer words.

Keywords: Data Science, Curriculum, keyword 3

1 Introduction

The need for training in data science is clear. An estimated 11.5 million new data science jobs are projected to be created by 2026, while employment of data scientists is projected to grow by 36 percent from 2021 to 2031 (U.S. Bureau of Labor Statistics, 2023). In response, college and university faculty are starting to develop and implement data science courses (Schwab-McCoy, Baker, & Gasper, 2020). The purpose of this paper is to continue to promote conversations and provide recommendations around the development and implementation of data science courses through the lens of a modernized introductory data science course taught at Duke University.

This course is designed for students with little to no statistics, data science, or coding experience, a common hurdle identified by faculty when trying to adopt data science (Schwab-McCoy, Baker, & Gasper, 2020). By the end of this course, students are expected and able to clean, investigate, and communicate with data in a reproduceable manner while answering a targeted research question. Detailed learning objectives of this course include learning to explore, visualize, and analyze data in a reproducible and shareable manner through the use of R-studio and GitHub (cite). Through these programs, students gain experience in data wrangling and munging, exploratory data analysis, predictive modeling, and data visualization. These experiences are generated through a Prepare, Practice, and Perform learning model emersed in real-world questions and data.

To continue and progress the conversation around data science courses, we first examine current curriculum recommendations for undergraduate programs in data science, discuss the current state in undergraduate introductory data science curricula, and review current pedagogical recommendations on how to teach such courses. Next, we discuss the creation and implementation of curricular and pedagogical decisions made in designing the introductory data science course at Duke University. This includes detailing the Prepare, Practice, and Perform format, to support a large class of students with a diverse background in statistics, data science, and coding experience. Within this format, we provide examples of and describe activities and assessments given both in and outside of class. We extend discussions and provide recommendations for implementing and integrating computing tools, such as R-studio and GitHub, in a data science course through our experiences in our course. Lastly, we discuss challenges, and provide insight to help faculty wanting to adopt or adapt a course similar to introductory to data science at Duke University.

2 Data Science: A Review

Although the definition of data science is fluid, it can be generally defined the process transforing raw data into understanding, insight, and knowledge (Wickham & Grolemund). In practice, data scientists often describe their practice as a means to "gain insights" or "extract meaning" from data (Hernán … , 2019). In the following sections, we describe the current curricular and pedagogical recommendations for designing a course in data science.

2.1 Curriculum (what is being taught)

Until recently, data modeling made up the majority of data science curricula through classes housed in statistics and mathematics departments (Donoho, 2017). However, there has been a call to create more well-rounded curricula that better align students of with our working definition of data science. In 2017, Curriculum Guidelines for Undergraduate Programs in Data Science provided six major recommendations as to what practitioners of data science should be competent in: Computational and statistical thinking; Mathematical foundations; Model building and assessment; Algorithms and software foundation; Data curation; Knowledge transference—communication and responsibility (Veaux, et al., 2017). These guiding pillars offer an approach for students to develop into problem solvers who interact with, investigate, and make meaning with data. Other visions of data science include more broad divisions of activity: The activities of GDS are classified into six divisions: Data Gathering, Preparation, and Exploration; Data Representation and Transformation; Computing with Data; Data Modeling; Data Visualization and Presentation; and Science about Data Science (Donoh, 2017).

The Association for Computing Machinery (ACM) Education Council is a task force to explore and expand discipline-specific conversations around the field of data science (cite). This council acknowledges that data science curricula can be flexible, but should emphasize suggests that data science curricula should include applications designed towards building skills in computing, statistics, machine learning and mathematics.

Among existing curricular recommendations, there are computational and technological recommendations for data science classrooms. This includes the use of GitHub to ensure the concept of reproducibility and incorporating quantitative programming environments (such as R) (Donoho, 2017; Beckman et al., 2021). Among all recommendations, it is critical to immerse students in real-world data and provide open-ended projects. This allows students "early exposure to and experience with the full data science cycle", in a practical real-world context (Lui & Huang, 2017; CETINKAYA-RUNDEL, et al., 2022, pg. 3)

2.2 Pedagogical (how to teach)

Research has shown that the adoption, integration, and implementation of active learning teaching techniques can help promote student learning, achievement, and confidence in mathematics (e.g., Freeman et al., 2004).

References