OVERCOMING THE CHALLENGES IN DEVELOPING AN INTRODUCTORY COURSE ON THE FOUNDATIONS OF DATA SCIENCE

Jessica Jaynes and Sam Behseta
California State University Fullerton, USA
jjaynes@fullerton.edu

THE SOCAL DATA SCIENCE PROGRAM

With the increasing demand for data-related skills in the workforce, it is essential that data science courses for undergraduates, particularly those with no or limited statistical or programming background are developed and accessible for all scientific disciplines. Specifically, as a data scientist ranks among the top emerging jobs, it is critical that institutions of higher learning address this. Moreover, it has been shown that creating a social community to support students can improve retention and mitigate some challenges faced by underrepresented minority students in STEM.

The SoCal Data Science Program, funded by the National Science Foundation, brings together three institutions from California's higher education system, namely University of California, spearheading research and discovery, California State University, combining research and pedagogy, and Community College, offering two-year preparatory programs. Particularly, California State University Fullerton (CSUF), University of California Irvine, and Cypress College have joined together to create new roadmaps for recruiting, training, and mentoring a diverse generation of data scientists who are ready to join the workforce upon graduation.

FOUNDATIONS OF DATA SCIENCE COURSE

At CSUF in Spring 2022, a new lower-division course on the Foundations of Data Science was developed as a byproduct of the SoCal Data Science Program. There are various aspects of this course that make it unique, but particularly noteworthy is the light prerequisite of only Precalculus as well as no required programming experience. This course brings together components of mathematics, statistics, and computer science, all within the context of real-world applications. Topics include data wrangling, data summarization and visualization, introduction to linear predictive modeling, cross-validation for model validation, and binary classification. Throughout the course students learn these topics through hands on data analysis using RStudio where they are exposed to the foundations of *ggplot2* and *tidyverse*. Ultimately, the course concludes with a final project from an original real-world dataset.

The students enrolled in this course come from a variety of backgrounds and majors such as Mathematics, Computer Science, Biology, Public Health, Computer Engineering, and even Humanities. Given the diverse backgrounds of the students, there have been numerous challenges in developing an appropriate level of curriculum. To facilitate the community among the students, as well as engage students in peer mentoring, students are grouped together at the beginning of the semester and provided the opportunity to work together throughout the semester on various assignments both in and outside of the classroom. In just the two semesters that this course has been implemented there have been profound impacts on the students, including students changing their major to data science related field such as statistics, participation in undergraduate research experiences related to data science, as well as pursuing graduate school related to data science. As such, it is imperative that similar courses in data science continue to be developed at various types of institutions to provide students with the opportunity and to inspire them to explore the growing field of data science.

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. https://iase-web.org