Foundations of Data Science with Capstone

SMU, Summer 2023

Instructor: Dr. Jeho Park, Claremont McKenna College

Course Description (강의 개요)

Data science is the interdisciplinary and practical fields of study about the tools and theory behind using data to extract knowledge. It combines ideas from statistics, computer science, and particular domains in the physical and social sciences in order to make data-driven predictions and optimal decisions.

In recent years, the demand for data scientists, data analysts, and data engineers has been consistently high due to the increasing importance of data-driven decision making in various industries. With the rapid growth of big data and the need to extract insights from large datasets, organizations are seeking professionals with expertise in these fields.

This summer intensive course is designed with a practical approach. The students will be guided to learn the basics and move on to more advanced topics in 2 weeks. The final 1 week will be dedicated to practicing and presenting data science skills and competency through capstone projects.

데이터 과학은 데이터를 활용하여 지식을 추출하기 위한 도구와 이론에 관한 다학제적(interdisciplinary) 이며 실용적인 학문입니다. 데이터 과학은 통계학, 컴퓨터 과학, 물리학과 사회과학의 특정 분야에서 여러 아이디어를 결합하여 데이터에 근거한 예측과 최적의 결정을 내리는 데 사용됩니다.

최근 몇 년 동안 다양한 산업에서 데이터 기반의 의사 결정이 점점 중요해지면서 데이터 과학자, 데이터 분석가 및 데이터 엔지니어에 대한 수요가 꾸준히 높아지고 있습니다. 대용량 데이터의 급속한 증가와 대규모 데이터셋에서 통찰력을 추출해야 하는 요구에 따라 다양한 종류의 비즈니스와 정부 기관들은 이러한 분야의 전문 지식을 가진 전문가를 찾고 있습니다.

이 여름학기 집중과정은 실용적인 접근 방식으로 설계되었습니다. 학생들은 2 주 동안 기초를 학습하고 보다 고급 주제로 나아갈 수 있도록 마지막 1 주에 캡스톤 프로젝트를 진행하게 됩니다. 이 캡스톤 프로젝트를 통해 학생들은 데이터 프로젝트에서 일반적으로 쓰이는 기술과 역량을 실습하고 마지막에 프로젝트의 결과에 대한 발표를 하게 됩니다.

Course Goal: After taking this course, you'll be able to utilize the tools required to tackle a wide variety of data science challenges using R programming language. 이 과정을 이수한 후에는 R 프로그래밍 언어를 사용하여 다양한 데이터 과학적 도전에 대처하는 데 필요한 도구들을 활용할 수 있게 될 것입니다.

Session 1: Building Foundations (2 weeks): During the first 2 weeks, you will learn the foundations of data science including the basics of how to structure, visualize, transform, and model data. The primary programming language that we will be using is R, which is both simple to use and was designed around using data. The development environment we will be using is RStudio. Both R and RStudio are open source, and so may be downloaded to your personal laptop for free.

첫 2 주 동안은 데이터 과학의 기초를 학습할 것입니다. 데이터의 구조화, 시각화, 변환, 모델링 등에 대한 기본 개념을 배우게 됩니다. 주로 사용할 프로그래밍 언어는 데이터 활용을 중심으로 설계되어 사용하기 간편한 R 입니다. 개발 환경으로는 RStudio 를 사용할 것입니다. R 과 RStudio 는 모두 오픈 소스이며 개인 노트북 컴퓨터에 무료로 다운로드하고 인스톨할 수 있습니다.

Session 2: Practicing Knowledge (1 week): The Data Science Capstone is a team-based, project-based session, providing an opportunity to apply data science skills and knowledge obtained from the first 2 weeks of the foundation session. Teams of two or three students, under the direction of the course instructor and teaching assistants, will be working on a data project. The primary objective is to educate the students in solving real world data science problems in professional settings by leveraging their own computational, statistical, and domain skills. The capstone will end with a final presentation.

데이터 과학 캡스톤은 팀 기반의 프로젝트 세션으로, 기초 세션에서 얻은 데이터 과학 기술과 지식을 적용하는 기회를 제공합니다. 강사와 조교들의 지도 아래, 두 명 또는 세 명으로 구성된 팀이 데이터 프로젝트에 참여합니다. 주요 목표는 학생들을 전문적인 환경에서 실제 데이터 과학 문제를 해결할 수 있는 역량을 가진 컴퓨터사용, 통계, 및 각자의 도메인 기술을 활용하여 교육하는 것입니다. 캡스톤은 최종 발표로 마무리됩니다.

Student Expectation and Project Outcomes of Capstone

Students will work independently as well as collaboratively in the group for at least 3 to 4 hours a day on modeling, analysis, data manipulation/wrangling, visualization, etc. as needed for the project. At the same time, students will be guided to learn soft skills such as project management, presentation, and communication through the daily sync. Students will be expected to read necessary background articles/books, learn additional tools, and log their own work and time.

마지막 1 주간의 캡스톤 프로젝트 기간 동안에 학생들은 자기팀의 프로젝트에 필요한 모델링, 분석, 데이터 조작/가공, 시각화 등을 위해 하루에 최소 3~4 시간 독립적으로 및 팀원과 협업하여 작업할 것입니다. 동시에, 학생들은 매일 진행되는 짧은 씽크업미팅을 통해 프로젝트 관리, 발표, 의사 소통과 같은 소프트 스킬을 배우도록 교육 받을 것입니다. 학생들은 필요한 배경 자료나 관련된 책을 읽고, 추가 도구를 학습하며, 자신의 작업 시간과 작업 내용을 기록하며 프로젝트 관리에 대해 배울 것입니다.

Each team will provide the following outcomes in a timely manner:

- A project proposal detailing the intended goals, methodology, and deliverables
- A final presentation on the project outcomes
- A GitHub page containing all outcomes (README, data, codes, analyses results, etc.)

각 팀은 정해진 시간에 다음과 같은 결과물을 제출해야 합니다:

- 프로젝트 제안서: 목표, 방법론 및 성과물에 대한 구체적인 설명
- 프로젝트 발표: 프로젝트 결과에 대한 최종 발표
- GitHub 페이지: README, 데이터, 코드, 분석 결과 등 모든 결과물을 포함하는 깃허브 페이지

Textbook

We will be using R for Data Science by Hadley Wickham and Garrett Grolemund. This book is open source and can be found/downloaded online. Korean version is also available online.

Tools

- Programming Language: R
  - Download and install R from https://cran.rstudio.com/
- Integrated Development Environment: RStudio
  - Download and install RStudio from https://posit.co/download/rstudio-desktop/
- Program Management Tool: GitHub
  - Class GitHub repository will be provided.
- Learning Management System: Canvas
  - A Canvas online platform will be provided.

Daily Plan for Lecture, Lab, and Capstone

Date	Session	Hour	Title	Topics
7/17	Lecture 9 am – 12 pm	3	**Introduction and Visualization (1)**	Introduction to Data Science; RStudio; Graphical grammars (ggplot2)
	Lab 12 pm – 1 pm	1	Working in RStudio; First R Markdown; ggplot2 intro	Exercise 27.3.1, 27.4.7, 3.2.4, 3.3.1, 3.5.1
7/18	Lecture 9 am – 12 pm	3	**Data Visualization (2)**	Visualization in the tidyverse; Aesthetic mappings
	Lab 12 pm – 1 pm	1	More plotting methods	Exercise 3.5.1, 3.6.1, 3.8.1
7/19	Lecture 9 am – 12 pm	3	**Data Wrangling (1)**	R basics for data wrangling; Basic R objects and operators; Transforming data (dplyr)
	Lab 12 pm – 1 pm	1	Filtering data	Exercise 5.2.4
7/20	Lecture 9 am – 12 pm	3	**Data Wrangling (2)**	Transforming data (dplyr); filter, arrange, select, summarise, group_by; piping
	Lab 12 pm – 1 pm	1	Data transformation and piping	Exercise 5.3.1 #1, #3; 5.4.1 #2, #3, #4; 5.5.2; 5.6.7
7/21	Lecture 9 am – 12 pm	3	**Exploratory Data Analysis and Stats**	Why EDA; Some statistics; Variation; Covariation
	Lab 12 pm – 1 pm	1	EDA exercise	
7/24	Lecture 9 am – 12 pm	3	**Working with different data types**	Vectors; Factors; Strings; Dates and Times
	Lab 12 pm – 1 pm	1	Strings and Factors exercise	
7/25	Lecture 9 am – 12 pm	3	**Programming in R**	Pipes; Functions; Iteration and vectorization
	Lab 12 pm – 1 pm	1	Dates, Times, and Functions in R	
7/26	Lecture 9 am – 12 pm	3	**Modeling**	Regression; Linear models; Understanding residual
	Lab 12 pm – 1 pm	1	Modeling and visualizing data; capstone prep	
7/27	Capstone 9 am – 10 am	1	**Project kick-off**	Introduction and expectations; Project Proposal; Minimum Viable Product
7/31	Capstone 9 am – 11 am	2	**Final Presentation**	
	Total hours:	35		

(The schedule may subject to change depending on the speaker availability and other situations.)