Journal of The Korean Association
of Information Education Vol. 25, No. 3,
June 2021, pp. 459-469

http://dx.doi.org/10.14352/jkaie.2021.25.3.459

교사교육을 위한 공공 빅데이터 수집 및 스프레드시트 활용 기초 데이터과학 교육 사례 연구

허경

경인교육대학교 컴퓨터교육과

요약

본 논문에서는 현장 교사 및 예비교사를 위한 기초 데이터과학 실습 교육 사례를 연구하였다. 본 논문에서는 기초 데이터과학 교육을 위해, 스프레드시트 SW를 데이터 수집 및 분석 도구로 사용하였다. 이후 데이터 가공, 예측 가설 및 예측 모델 검증을 위한 통계학을 교육하였다. 또한, 수천명 단위의 공공 빅데이터를 수집 및 가공하고, 모집단 예측 가설 및 예측 모델을 검증하는 교육 사례를 제안하였다. 이와 같은 데이터과학의 기초 교육 내용을 담아, 스프레드시트 도구를 활용한 34시간 17주 교육 과정을 제시하였다. 데이터 수집, 가공 및 분석을 위한 도구로서, 스프레드시트는 파이썬과 달리, 프로그래밍 언어 및 자료구조에 대한 학습 부담이 없고, 질적 데이터와 양적 데이터에 대한 가공 및 분석 이론을 시각적으로 습득할 수 있는 장점이 있다. 본 교육 사례 연구의 결과물로서, 세가지 예측 가설 검증 사례들을 제시하고 분석하였다. 첫 번째로, 양적 공공데이터를 수집하여 모집단의 그룹별 평균값 차이 예측 가설을 검증하였다. 두 번째로, 질적 공공데이터를 수집하여 모집단의 질적 데이터 내 연관성 예측 가설을 검증하였다. 세 번째로, 양적 공공데이터를 수집하여 모집단의 양적 데이터 내 상관성 예측 가설 검증에 따른 회귀 예측 모델을 검증하였다. 그리고 본 연구에서 제안한 교육 사례의 효과성을 검증하기 위해, 예비교사와 현장교사의 만족도분석을 실시하였다.

키워드 : 공공데이터, 데이터과학, 스프레드시트, 양적데이터, 질적데이터, 데이터예측

A Case Study of Basic Data Science Education using Public Big Data Collection and Spreadsheets for Teacher Education

Kyeong Hur

Dept. of Computer Education, Gyeongin National University of Education

Abstract

In this paper, a case study of basic data science practice education for field teachers and pre-service teachers was studied. In this paper, for basic data science education, spreadsheet software was used as a data collection and analysis tool. After that, we trained on statistics for data processing, predictive hypothesis, and predictive model verification. In addition, an educational case for collecting and processing thousands of public big data and verifying the population prediction hypothesis and prediction model was proposed. A 34-hour, 17-week curriculum using a spreadsheet tool was presented with the contents of such basic education in data science. As a tool for data collection, processing, and analysis, unlike Python, spreadsheets do not have the burden of learning programming languages and data structures, and have the advantage of visually learning theories of processing and analysis of qualitative and quantitative data. As a result of this educational case study, three predictive hypothesis test cases were presented and analyzed. First, quantitative public data were collected to verify the hypothesis of predicting the difference in the mean value for each group of the population. Second, by collecting qualitative public data, the hypothesis of predicting the association within the qualitative data of the population was verified. Third, by collecting quantitative public data, the regression prediction model was verified according to the hypothesis of correlation prediction within the quantitative data of the population. And through the satisfaction analysis of pre-service and field teachers, the effectiveness of this education case in data science education was analyzed.

Keywords : public data, data science, spreadsheet, quantitative data, qualitative data, data prediction

논문투고 : 2021-05-19
논문심사 : 2021-05-25
심사완료 : 2021-05-31

1. 서론

1.1 연구의 필요성 및 목적

컴퓨터가 대량의 데이터를 수집, 가공 및 처리해서 다양한 상황을 예측하는 기술을 데이터과학이라고 간주하고 있다. 이를 위해, 첫째로, 수집한 데이터의 속성을 이해하기 위해서는 질적 데이터인지, 양적 데이터인지 구분하고, 그 데이터가 어떠한 처리 과정을 거쳐 얻어진 것인지 이해할 수 있어야 한다. 두번째로, 수집한 데이터를 분석하려면, 그 데이터의 어떠한 속성들을 분석할 것인지 그 대상과 관계를 먼저 정의하고, 어떠한 데이터 가공 작업을 실시할 것인지 논리적으로 설계할 수 있어야 한다. 세 번째로, 데이터 가공 및 자동화 처리를 위해, 적절한 SW를 선택하여 데이터를 분석한다[1-5].

이러한 데이터과학 교육을 위해, 공공기관에서도 교육 실습용으로 사용할 수 있는 데이터를 공개하고 있다[6-9]. 데이터과학 교육을 통해 양성된 교사들은, 각 교육 기관에서 중요한 의사결정 시에 핵심적인 역할을 할 수 있도록, 관련된 많은 데이터를 축적하고 가공하여 다양한 분석을 통해, 사회적 현상을 예측할 수 있는 인사이트를 갖추어야 한다. 이것이 교사교육을 위한 데이터과학 교육의 목표가 되어야 한다.

데이터과학 교육에 대한 선행 연구를 살펴보면, '4차 산업혁명 시대 데이터과학 교육 방향성 모색' 이라는 주제로 데이터과학 교육을 통해 갖추어야 할 5가지 소양을 제시된 바 있고[10], 박윤수(2020)은 전문가 조사를 통해, 빅데이터 가공과정에 반드시 빅데이터 처리 프레임워크 또는 고성능 컴퓨터가 필요한 것은 아니며, 컴퓨터 과학적 지식과 스킬보다는 빅데이터 분석 방법과 응용 방법 중심으로 데이터과학 교육을 실시해야 한다고 제시하였다[11]. 이러한 데이터과학교육에 대한 요구를 바탕으로, 허경(2020)은 초등영역에 초점을 맞춘 데이터과학 교육 사례를 제안하였고, 이를 위해, 엔트리에서 제공하는 공공 스몰 데이터를 사용한 데이터 변수 평균값 비교 가설 검증 사례와 데이터 변수 간 상관관계 분석 가설 검증 사례를 초등 데이터과학 교육 단계에 따라 제안하였다[12]. 홍지연(2020)은 초등학생의 데이터 리터러시를 키워줄 수 있는 AI 데이터과학 교육 프로그램을 개발하고, 데이터 리터러시의 데이터 이해, 수집, 분석, 표현 역량 향상에 있어 효과성을 검증하였다[13]. 그리고 김봉철(2021)은 초등학생이 마이크로비트를 활용하여 데이터를 수집하는 프로그램을 제작해 보고, 수집된 데이터를 분석하여 결과를 도출하는 데이터 과학의 단계를 수행하는 교육프로그램을 개발하고, 컴퓨팅 사고력 향상 효과를 검증하였다[14].

이와 같이, 초중등 교육을 위한 데이터과학교육 관련 연구는 활발하게 진행되는 데 비해, 예비교사와 현직교사를 위한 데이터과학교육 관련 연구는 상대적으로 많이 진행되지 못하였다. 현직교사를 위한 데이터과학교육 관련 선행 연구로, 구덕회(2020)은 교육전문대학원 AI교육을 위한 데이터과학 교육프로그램을 제시하였다[15]. 제시한 데이터과학 교육프로그램은 AI교육대학원에 재학 중인 현장교사를 대상으로 한다. 여기서 제시된 15주 커리큘럼은 파이썬을 이용한 Anaconda 또는 구글 Colab 플랫폼을 이용하여, 프로그래밍 활동 중심으로 데이터 분석하는 고급 과정이다. 이 고급 교육과정은 파이썬 실행창 결과화면을 보면서, 데이터 분석과정을 이해해야 한다. 본 교육과정의 단점은 자료구조 프로그래밍 역량과 통계학적 지식이 없는 경우, 데이터 저장, 가공 및 분석 과정을 시각적으로 이해하기 어렵다는 것이며, 결국, 질적 데이터와 양적데이터에 대한 분석 이론 등 통계학 교육을 별도로 실시해야 한다는 것이다.

이러한 단점을 보완하기 위해, 본 논문에서는 예비교사와 현직교사를 대상으로 파이썬을 이용한 고급 데이터과학 교육을 실시하기 전에 필요한 기초 데이터과학 실습 교육 사례를 연구하였다. 이를 통해, 스프레드시트 도구를 활용한 34시간 17주 교육 과정을 제시하였다. 데이터 수집, 가공 및 분석을 위한 도구로서, 스프레드시트는 파이썬과 달리, 프로그래밍 언어 및 자료구조에 대한 학습 부담이 없고, 질적 데이터와 양적 데이터에 대한 가공 및 분석 이론을 시각적으로 습득할 수 있는 장점이 있다. 제안한 34시간 17주 교육 과정은 데이터 가공, 예측 가설 및 예측 모델 검증을 위한 통계학을 교육하고, 수천명 단위의 공공 빅데이터를 수집 및 가공하고, 모집단 예측 가설 및 예측 모델을 검증하는 실습교육을 포함한다.

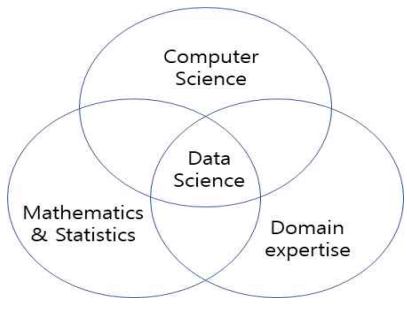
본 교육 사례 연구의 결과물로서, 세가지 예측 가설 검증 사례들을 제시하고 분석하였다. 첫 번째로, 양적 공공데이터를 수집하여 모집단의 그룹별 평균값 차이

예측 가설을 검증하였다. 두 번째로, 질적 공공데이터를 수집하여 모집단의 질적 데이터 내 연관성 예측 가설을 검증하였다. 세 번째로, 양적 공공데이터를 수집하여 모집단의 양적 데이터 내 상관성 예측 가설 검증에 따른 회귀 예측 모델을 검증하였다. 그리고 본 연구에서 제안한 교육 사례의 효과성을 검증하기 위해, 예비교사와 현장교사의 만족도분석을 실시하였다.

2. 선행 연구 분석

2.1 데이터과학의 개념과 스프레드시트의 데이터 분석 기능

(Fig. 1)에 데이터과학자가 갖추어야 할 3가지 역량을 제시하였다. 데이터과학자는 데이터 수집, 가공 및 분석을 통해 인사이트를 추출하여 더 나은 의사결정에 도움을 주는 전문가이다. 데이터과학자의 역할은 컴퓨터 과학을 바탕으로 데이터 수집 및 분석 인프라와 플랫폼 구축한다. 그 후, 수학 및 통계학 지식을 바탕으로 데이터를 가공 및 분석하고, 해당 도메인 영역에 대한 전문적 지식을 바탕으로 모집단의 데이터 예측 가설 및 모델을 검증한다[3-4]. 스프레드시트의 장점은 데이터 중심 사고를 교육하는 데 효과적이고, 쉽고 다양한 기능, 범용성을 갖는다는 것이다. 온라인 설문조사의 데이터 수집 결과는 스프레드시트 파일로 저장된다. 데이터 집계는 스프레드시트의 필터링, 피벗테이블, 슬라이서 기능을 사용한다. 데이터분석은 스프레드시트의 분석도구를 사용하고, 데이터 시각화는 스프레드시트의 조건부셀 서식, 차트, 파워피벗 기능을 사용한다[5].



(Fig. 1) Components of Data Science

2.2 파이썬 데이터분석 플랫폼을 이용한 고급 데이터과학 교육과정

구덕회(2020)는 교육전문대학원 AI교육을 위한 데이터과학 교육프로그램을 <Table 1>과 같이 제시하였다[15]. 여기서 제시된 15주 커리큘럼에서는 2주차와 3주차에서 데이터과학의 기반이 되는 기초 통계학 내용을 강의한다. 4주차부터 7주차까지는 파이썬에서 자료구조를 생성하여 데이터를 저장하고 가공하는 기법을 교육한다. 8주차와 9주차는 파이썬에서 데이터를 시각화하여 표현하는 기법을 교육한다. 10주차부터 13주차까지 머신러닝을 이용하여 회귀, 분류 및 군집에 따른 데이터 분석결과를 실습한다. 14주차에서는 머신러닝을 적용하여 도출된 예측모델을 검증하는 방법을 교육하고, 15주차에서 공공데이터를 활용한 데이터 분석 실습을 실시한다. 본 15주 교육과정은 파이썬 언어 기반 Anaconda 또는 구글 Colab 플랫폼을 사용하여 데이터를 분석하는 고급 과정이다. 본 교육 과정의 단점은 파이썬 자료구조 프로그래밍을 통해, 데이터를 분석한다는 것이다. 이는 데이터 중심 사고 과정, 즉, 데이터 저장, 가공 및 분석 과정을 시각적으로 이해하기 어렵다는 것을 의미한다. 그리고 질적 데이터와 양적 데이터 자체 특성을 이해하는 기초 교육 내용과 모집단에 대한 예측 모델을 검증하는 통계학 교육 내용이 다소 부족하다.

<Table 1> Advanced data science curriculum using the Python data analysis platform

Unit	Activity Contents
1	[Introduction of Data Science] - definition and purpose, examples, big data
2	[Data statistics] - maximum & minimum values, mean, standard deviation, - probability, hypothesis testing, linear algebra
3	[Statistics & Probability] - central limit theorem, z-score, probability
4	[Table type data processing] - python array type data, array indexing, array slicing, dimension
5	[Python data structure & arithmetic operation] - pandas, series, data frame, indexing, selection, filtering
6	[Data preprocessing] - handling missing data, remove duplication, data transformation

7	[Data frame & function] - data content check, check summary information, statistical function
8	[Basic data visualization] - matplotlib, line graph, bar graph, scatter plot
9	[Advanced data visualization] - seaborn, folium, hisogram, grid
10	[Machine learning] - supervised and unsupervised learning, classification, regression, clustering, dimension reduction
11	[Regression analysis] - simple regression analysis, polynomial regression analysis, multiple regression analysis
12	[Classification] - KNN, SVM, Decision Tree
13	[Clustering] - k-Means, DBSCAN
14	[Model validation] - overfitting, underfitting, parameter tuning, cross-validation
15	[Practice] - practice using public data

3. 스프레드시트 데이터분석 도구를 활용한 34시간 17주 기초 데이터과학 교육과정

기존 데이터과학 교육과정 연구결과[13]의 단점을 보완하기 위해, 본 논문에서는 파이썬을 이용한 고급 데이터과학 교육[13]을 실시하기 전에 필요한 기초 데이터과학 실습 교육 사례를 연구하였고, 연구결과로 스프레드시트 도구를 활용한 34시간 17주 교육 과정을 제시하였다. 데이터 저장(수집), 가공 및 분석을 위한 도구로서, 스프레드시트는 파이썬과 달리, 프로그래밍 언어 및 자료구조에 대한 학습 부담이 없고, 질적 데이터와 양적 데이터에 대한 저장, 가공 및 분석 이론을 시각적으로 학습할 수 있는 장점이 있다. 제안한 34시간 17주 교육과정은 데이터 특성 이해를 위한 기초 데이터 교육을 포함하고, 데이터 대푯값 추출, 예측 가설 및 예측 모델 검증을 위한 통계학을 교육한다. 또한, 수천명 단위의 공공 빅데이터를 수집 및 가공하고, 모집단 예측 가설 및 예측 모델을 검증하는 실습교육을 포함한다.

<Table 2>는 스프레드시트 데이터 분석 도구를 활용한 34시간 17주 교육 과정을 나타낸다. <Table 2>의 교육과정은 참고문헌[5]에 기초하여 확대 개발되었다. 1주차에서는 데이터과학에 의한 문제해결과정을 교육한다. 이를 위해, 세부적으로 데이터의 종류, 통계적 방법과 데이터 변수, 스프레드시트를 이용한 데이터 과학을 강의한다. 2주차에서는 설문 조사표를 통해 데이터를 만드는 방법, 데이터를 수집하는 방법 그리고 데이터를 변환하는 방법을 각각 강의한다. 3주차에서는 표본과 모집단 모수, 데이터와 척도, 데이터의 분포와 같은 기초적인 데이터 해석 방법을 강의한다. 4주차에서는 데이터 입력, 질적 데이터 집계 및 양적 데이터 집계 방법을 각각 강의하고 실습한다. 5주차에서는 수집한 질적 데이터와 양적 데이터 내 속성 간 교차 집계 방법으로, 교차표 작성 방법과 피벗 테이블 만드는 방법을 각각 강의하고 실습한다. 6주차에서는 데이터 시각화 주제로, 통계 그래프의 종류와 특징, 통계 그래프 작성의 기초를 각각 강의하고 실습한다. 7주차에서는 수집한 데이터 그룹의 위치 특성을 나타내는 통계 대푯값으로 평균값, 중앙값 및 최빈값 등을 강의하고 실습한다. 8주차에서는 수집한 데이터 그룹의 크기를 나타내는 통계 대푯값으로 사분위수, 분산, 표준편차, 산포도 그리고 변화율을 강의하고, 스프레드시트의 데이터 분석 도구를 사용하여 수집한 데이터 그룹의 위치와 크기 대푯값을 구하는 방법을 실습한다.

<Table 2> Basic data science curriculum using spreadsheets

Unit	Activity Contents
1	[Data Science] - Data science using a variety of data - statistical methods and variables, spreadsheets
2	[Data collection] - Data generation, data collection, data conversion method
3	[Data Processing] - Samples and population parameters - data and scale, distribution of data
4	[Data aggregation] - Data input, qualitative data aggregation, quantitative data aggregation
5	[Cross-aggregation] - How to create a crosstab, - How to create a pivot table, cross-aggregate in a survey
6	[Statistics graph] - Types and characteristics of statistical graphs, basics of creating statistical graphs
7	[Basic statistical value practice] - Comparison of representative value (position

	parameter), average value, median value, and mode of magnitude
8	[Basic statistical value practice] - Comparing standard deviations, rate of change, using analysis tools
9	[Guessing and judgment] - Population mean/ variance/ratio estimation, test order
10	[Mean value difference prediction test] - Test for mean difference in populations when there are three or more groups
11	[Mean value difference prediction test] - A lecture on a report on mean value difference prediction test in quantitative data through collection of public big data
12	[Qualitative data analysis] - Relevance measurement: an indicator that measures how relevant the population is
13	[Qualitative data analysis] - Determining Relevance: How to find out if there is a relation in a population
14	[Qualitative data analysis] - A lecture on a report on correlation analysis in qualitative data through collection of public big data
15	[Quantitative data analysis] - Correlation coefficient that measures the degree of correlation
16	[Quantitative data analysis-machine learning] - Single regression and multiple regression model test analysis predicting from quantitative data
17	[Quantitative data analysis-machine learning] - A lecture on regression model test analysis report in quantitative data through public big data collection

본 기초 데이터과학 교육과정의 1주차에서 8주차까지는 질적 데이터와 양적 데이터 자체 특성을 해석 및 이해하는 기초 교육을 포함한 데이터 중심 사고 활동을 시각적으로 실시하며 교육한다. 즉, 스프레드시트를 사용하여 수집한 데이터가 한눈에 모두 저장되어 보이고, 질적 데이터와 양적 데이터를 보면서 직접 수정한다. 또한, 교차 집계한 데이터 속성 간의 관계를 한눈에 보며, 분석에 필요한 데이터 가공 과정을 학습한다. 그리고 가공된 데이터의 위치와 크기 대푯값을 표현하고 해석하는 방법을 학습한다.

9주차에서는 표본으로부터 모집단의 평균값/분산/비율 추정하는 방법을 각각 강의하고 실습한다. 10주차에서는 표본으로부터 모집단 내 그룹 간 평균값 차이를 검정하는 TTEST와 동일 표본의 사전 사후 검정에 따른 모집단의 평균값 차이를 검정하는 TTEST를 각각 강의하고 스프레드시트의 데이터 분석 도구를 사용하여 실습한다. 11주차에서는 공공 빅데이터를 수집하여, 양적 표본 데이터로부터 모집단 내 그룹간 평균값 차이를 검정하는 TTEST 보고서 작성 사례를 강의하고 실습한다.

12주차에서는 질적 자료에서 연관성 측정 방법을 강의하고 실습한다. 13주차에서는 질적 자료에서 연관성 판단 방법을 강의하고 실습한다. 14주차에서는 공공 빅데이터 수집을 통한 질적 데이터 내 속성 간 연관성 분석 보고서 작성 사례를 강의하고 실습한다. 15주차에서는 양적데이터 내 속성 간 상관성을 측정하는 방법을 강의하고 실습한다. 16주차에서는 양적 데이터 표본으로부터 모집단 데이터 속성을 예측하는 머신러닝 단일/다중 회귀 모델을 디자인하고 검증하는 방법을 강의하고 실습한다. 17주차에서는 공공 빅데이터 수집을 통한 양적 데이터 표본으로부터 모집단 데이터 내 속성 간 머신러닝 단일/다중 회귀 모델 디자인 및 검증 보고서 작성 사례를 강의하고 실습한다.

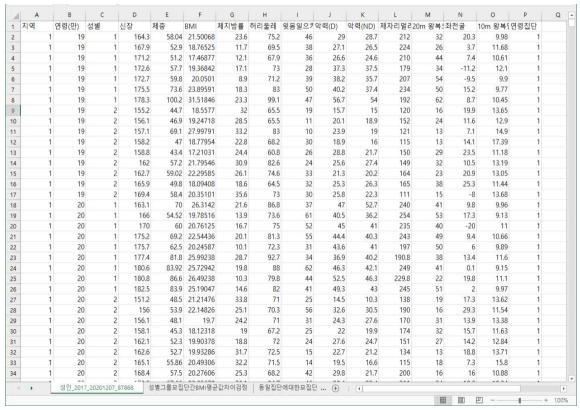
9주차부터 17주차까지 교육내용의 핵심은 수집한 질적/양적 표본 데이터를 가공한 후, 모집단의 특성을 나타내는 대표값을 통계적으로 예측하는 가설을 설정하고 검정하는 것이다. 또한, 수집한 양적 표본 데이터로부터 모집단의 특정 속성값을 예측하는 모델을 만들고 검정하는 것이다. 본 교육 과정은 데이터 중심 사고 교육 과정으로, 수집한 표본 데이터와 모집단 데이터의 차이를 강조한다. 그리고 수집한 표본 데이터와 모집단 데이터 간의 관계를 예측하고 예측한 결과를 검정하는 통계학적 방법을 중점적으로 교육한다.

4. 공공 빅데이터 수집을 통한 모집단의 그룹별 평균값 차이 예측 분석 교육 사례

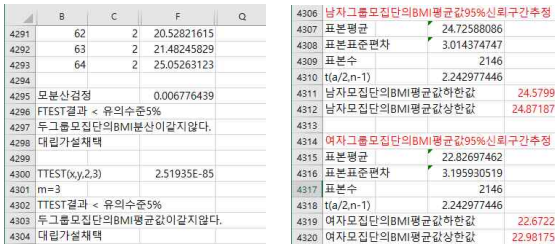
4장부터 6장에서 소개하는 데이터 분석 교육 사례는 마이크로데이터 통합서비스 (http://mdis.kostat.go.kr)에서 공공데이터를 수집하여 분석한 보고서 작성 교육 사례이다. 본 장에서는 첫 번째로, 양적 공공 빅데이터를 수집하여, 모집단의 그룹별 평균값 차이 예측 가설을 검정한 교육 사례를 제시하였다.

4.1 모집단 내 남자 여자 그룹 간 BMI 평균값 차이 예측 가설 검정 사례

(Fig. 2)는 2017년도 성인 4292명 국민체력측정 통계 표본 데이터를 수집한 결과이다. 4292명의 데이터를 성별 속성을 기준으로 정렬하여, 남성 2146명과 여성 2146명 두 그룹으로 분류하였다. 그리고 대한민국 성인 전체 모집단의 BMI 양적 데이터 속성에 있어, 남자와 여자 그룹 간 BMI 평균값이 차이가 있는 지 검정한 결과와 모집단 내 그룹별 BMI 평균값을 추정한 결과를 (Fig. 3)에 나타내었다.



(Fig. 2) Statistical data for measuring national physical fitness of 4292 adults in 2017



	B	C	F	Q
4291	62	2	20.52821615	
4292	63	2	21.48245829	
4293	64	2	25.05263123	
4294				
4295	모분산검정		0.006776439	
4296	FTEST결과 < 유의수준5%			
4297	두그룹모집단의BMI분산이같지않다.			
4298	대립가설채택			
4299				
4300	TTEST(x,y,2,3)		2.51935E-85	
4301	m=3			
4302	TTEST결과 < 유의수준5%			
4303	두그룹모집단의BMI평균값이같지않다.			
4304	대립가설채택			

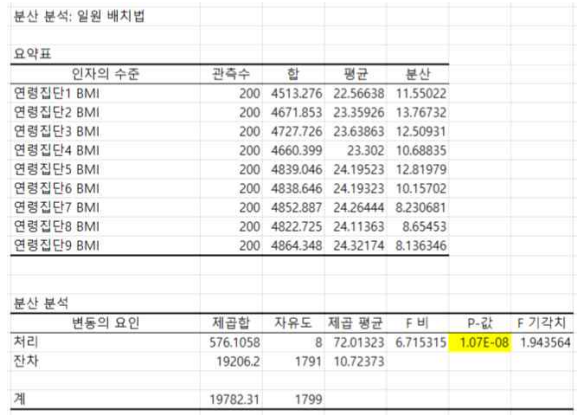
4306	남자그룹모집단의BMI평균값95%신뢰구간추정		
4307	표본평균	24.72588086	
4308	표본표준편차	3.014374747	
4309	표본수	2146	
4310	t(a/2,n-1)	2.242977446	
4311	남자모집단의BMI평균값하한값		24.5799
4312	남자모집단의BMI평균값상한값		24.87187
4313			
4314	여자그룹모집단의BMI평균값95%신뢰구간추정		
4315	표본평균	22.82697462	
4316	표본표준편차	3.195930519	
4317	표본수	2146	
4318	t(a/2,n-1)	2.242977446	
4319	여자모집단의BMI평균값하한값		22.6722
4320	여자모집단의BMI평균값상한값		22.98175

(Fig. 3) Results of testing the difference in the mean value of BMI between male and female groups in the entire population of Korean adults and estimating the mean value

모집단 내 그룹 간 평균값 차이를 검정하기에 앞서, 두 그룹간 모집단의 분산값 차이를 유의수준 5%에서 먼저 검정하였다. 모분산 검정 결과로 대립가설이 채택되어, 모집단 그룹 간 평균값 차이 TTEST 입력 변수값이 결정되었고, 유의수준 5% TTEST 검정 결과로 대한민국 성인 남자 여자 그룹 간 BMI 양적 데이터의 평균값은 차이가 있다는 대립 예측 가설이 유효한 것으로 분석하였다. 또한 t분포를 이용하여, 95% 신뢰구간으로 대한민국 성인 남자와 여자 그룹 BMI 평균값의 상한과 하한 범위를 각각 예측하였다.

4.2 모집단 내 다수 연령집단 그룹 간 BMI 평균값 차이 예측 가설 검정 사례

(Fig. 4)는 2017년도 성인 4292명 국민체력측정 통계 표본 데이터를 연령집단 속성에 따라 정렬한 후, 연령집단 9개 그룹별로 200개 BMI 표본만 추출하고 분리 저장한 1800개 BMI 데이터를 일원배치분산분석 도구로 분석한 결과를 나타낸다. 대한민국 성인 모집단 내 연령집단 9개 그룹 간 BMI 평균값 차이 검정 확률 P가 (Fig. 4)에서 유의수준 5%보다 작은 값을 나타내어 대립가설이 채택되었다. 이에 따라, 대한민국 성인 모집단 내 연령집단 9개 그룹 간 BMI 평균값은 서로 같지 않다는 대립 예측 가설이 유효한 것으로 분석하였다.



분산 분석: 일원 배치법

요약표

인자의 수준	관측수	합	평균	분산
연령집단1 BMI	200	4513.276	22.56638	11.55022
연령집단2 BMI	200	4671.853	23.35926	13.76732
연령집단3 BMI	200	4727.726	23.63863	12.50931
연령집단4 BMI	200	4660.399	23.302	10.68835
연령집단5 BMI	200	4839.046	24.19523	12.81979
연령집단6 BMI	200	4838.646	24.19323	10.15702
연령집단7 BMI	200	4852.887	24.26444	8.230681
연령집단8 BMI	200	4822.725	24.11363	8.65453
연령집단9 BMI	200	4864.348	24.32174	8.136346

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
처리	576.1058	8	72.01323	6.715315	1.07E-08	1.943564
잔차	19206.2	1791	10.72373			
계	19782.31	1799				

(Fig. 4) Results of testing the difference in mean values between 9 groups of Korean adult age groups

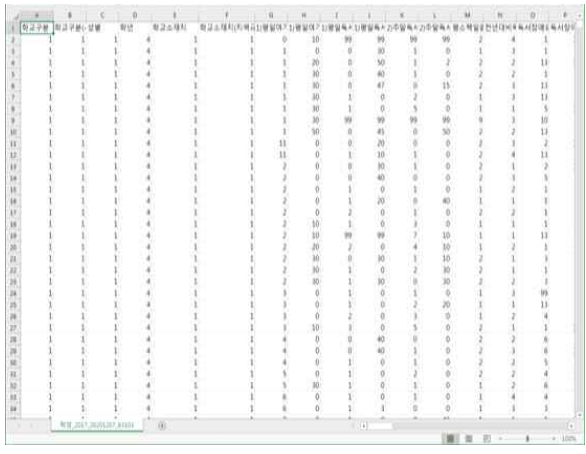
5. 공공 빅데이터 수집을 통한 질적 데이터의 연관성 예측 분석 교육 사례

본 장에서는 마이크로데이터 통합서비스 웹사이트

(http://mdis.kostat.go.kr)에서 수집한 공공 질적 표본 데이터로부터 모집단 데이터 내 속성 간 연관성을 분석하여 의사결정에 활용한 교육 사례를 제시한다.

5.1 본 공공데이터를 수집한 목적 정의

(Fig. 5)와 같이 2017년도 초중고학생 3107명 독서관련 공공데이터를 수집하여, 학생 연령과 관련된 '1번 학교 급'과 '13번 평소책읽는빈도' 속성 간의 질적 데이터 연관성 여부와 연관성 강도를 조사하여, 모집단에 대한 독서 장려 관련 지원 정책을 결정할 때, 학교급을 고려할 필요가 있는지 알아보고자 한다. 그리고, '61번 학교 도서관이용회수'와 '13번 평소책읽는빈도' 속성 간의 질적 데이터 연관성 여부와 연관성 강도를 조사하여, 모집단에 대한 독서 장려 관련 지원 정책을 결정할 때, 학교 도서관 시설 및 이용 지원 정책을 고려할 필요가 있는지 알아보고자 한다.



(Fig. 5) Collecting public data related to reading of 3107 elementary, middle and high school students in 2017

5.2 질적 데이터 변수 간 모집단에서의 연관성 여부와 연관성 강도 분석 절차

서로 다른 2개 질적 데이터 속성 변수들을 지정하고, 지정한 변수들간 교차표를 사용하여 카이제곱 독립성 검정(또는 예이츠, 피셔 검정)을 유의수준 5%로 수행한 결과를 구하고 모집단에서 연관성을 분석한다. 그 후, 모집단에서 연관성이 있는 것으로 판정되었다면, 율의 Q 또는 크라머의 V값을 구하여 연관성의 강도를 구하고 분석한다. 단, 모집단에서 연관성이 없는 것으로 판정되었다면, 또 다른 2개 질적 데이터 속성 변수들을 지정하여 모집단에서 연관성을 갖는지 분석해보고, 모집단에서 연관성을 갖는 질적 데이터 속성 변수들로 제한하여, 의사결정과정에 활용하도록 한다.

5.3 질적 데이터 변수 간 모집단에서의 연관성 여부와 연관성 강도 분석 사례

학생 연령과 관련된 '1번 학교 급'과 '13번 평소책읽는빈도' 항목 간의 질적 데이터 연관성 여부와 연관성 강도를 조사하여, 모집단에 대한 독서 장려 관련 지원 정책을 결정할 때, 학교급을 고려할 필요가 있는지 알아보고자 한다. 피벗테이블 기능을 이용하여, (Fig. 6)과 같이, 지정한 변수들간 교차표를 만들어서 카이제곱 독립성 검정을 유의수준 a=0.05로 수행한 결과를 구한다. (Fig. 6)의 결과로부터, 카이제곱값 > X2, P<a 이므로 대립가설이 채택되고, 연령에 해당하는 학교구분과 평소 책 읽는 빈도는 모집단에서도 연관성이 있다는 것을 알 수 있다. 분류항목이 2개 이상이므로, 크라머의 V를 구한다. 그 결과, V=0.2818 로서, V값이 0.5보다 크지않아, 연령에 해당하는 학교구분과 평소 책 읽는 빈도는 모집단에서도 연관성이 크지 않은 것으로 분석되었다.

		평소 책 읽는 빈도					
	행 레이블	매일	일주일에몇번	한달에한두번	몇달에한번	전혀안읽는다	9 총합계
학교구분	초등학교	282	576	118	63	22	36 1097
	중학교	113	391	217	180	58	81 1040
	고등학교	94	293	292	296	112	105 1192
	총합계	489	1260	627	539	192	222 3329

무응답을 제외한 관측도수

		평소 책 읽는 빈도					
		매일	일주일에몇번	한달에한두번	몇달에한번	전혀안읽는다	총합계
학교구분	초등학교	282	576	118	63	22	1061
	중학교	113	391	217	180	58	959
	고등학교	94	293	292	296	112	1087
	총합계	489	1260	627	539	192	3107

무응답을 제외한 기대도수

		평소 책 읽는 빈도					
		매일	일주일에몇번	한달에한두번	몇달에한번	전혀안읽는다	총합계
학교구분	초등학교	167.0	430.3	214.1	184.1	65.6	1061
	중학교	150.9	388.9	193.5	166.4	59.3	959
	고등학교	171.1	440.8	219.4	188.6	67.2	1087
	총합계	489	1260	627	539	192	3107

(Fig. 6) A case of analysis of relationship strength and association between qualitative data variables in the population

5.4 모집단에서 연관성을 갖는 질적 데이터 변수간의 분석 결과를 의사결정에 활용한 사례

초중고학생 3107명 무기명 표본 공공데이터를 분석하였다. 여기서 연령에 해당하는 학교구분과 평소 책 읽는 빈도 속성 간 질적 데이터 연관성을 분석한 결과로부터, 모집단에 대한 청소년 독서 장려 정책을 펼칠 때, 연령과 관련된 학교 급에 따라 차별화된 지원 정책을 펼칠 필요가 크지 않다는 결론을 내릴 수 있다.

6. 공공 빅데이터 수집을 통한 양적 데이터의 상관성 및 회귀 모델 예측 분석 교육 사례

대부분의 데이터 센터 사이트에서 질적 데이터를 수집하는 것은 비교적 용이하다. 그러나, 양적 데이터를 수집하는 것은 검색 시간이 소요된다. 공개된 데이터들은 전문용어가 많고, 부분적인 데이터들이 많아 데이터 결합 및 가공이 필수적인 경우가 많다. 본 장에서는 마이크로데이터 통합서비스 (http://mdis.kostat.go.kr)에서 수집한 공공 양적 표본 데이터로부터 모집단 데이터 내 속성 간 상관성과 회귀 모델을 검정 및 분석하여 의사결정에 활용한 교육 사례를 제시한다.

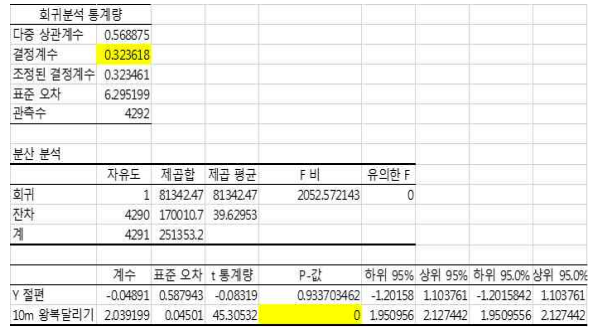
6.1 본 공공데이터를 수집한 목적 정의

(Fig. 2)에 제시한 2017년도 성인 4292명 국민체력측정 통계결과로부터, 비만도 및 체지방률과 같은 주요 건강 지표와 각종 기록 간의 관계를 분석하고자 하였다. 이를 통해, 체지방률과 20m 왕복달리기와 같은 기록 간의 관계들을 도출하여 국민들의 야외 운동을 장려하고자 하였다. 모집단에서 개인의 건강 지표와 각종 운동 종목 기록 간의 관계를 도출하면, 국민 개개인들에게 건강에 대한 관심을 크게 높일 수 있다.

6.2 양적 데이터 변수 간 모집단에서의 상관성 여부와 회귀 모델 예측 분석 절차

(Fig. 7) Correlation matrix between quantitative data variables in collected public data

(Fig. 7)과 같이 수집한 공공데이터 파일에서 3개 이상 양적 데이터 속성 변수들을 지정하고, 지정한 변수들 간 상관행렬을 구한다. 지정한 변수들 중 하나를 독립변수(설명변수)로 정하고, 하나를 종속변수(피설명변수)로 정한다. 데이터 분석 메뉴를 이용하여, 앞에서 지정한 변수들 간의 머신러닝 단순회귀모델 예측식을 구한다. 그리고 본 독립변수가 모집단에서 유효한 변수인지 0.05 유의수준으로 검정하고, 결정계수 값을 통해 예측값이 적합했는지 분석한다. 여기서는 모집단에서 유효한 독립변수와 종속변수 간의 회귀분석결과만이 유효한 분석결과로 정의한다.



회귀분석 통계량	
다중 상관계수	0.568875
결정계수	0.323618
조정된 결정계수	0.323461
표준 오차	6.295199
관측수	4292

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	81342.47	81342.47	2052.572143	0
잔차	4290	170010.7	39.62953		
계	4291	251353.2			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	-0.04891	0.587943	-0.08319	0.933703462	-1.20158	1.103761	-1.2015842	1.103761
10m 왕복달리기	2.039199	0.04501	45.30532	0	1.950956	2.127442	1.9509556	2.127442

(Fig. 8) Case of simple regression analysis between valid quantitative data variables (dependent-independent) in the population

6.3 양적 데이터 변수 간 모집단에서의 상관성 여부와 단순 회귀 모델 예측 분석 사례

(Fig. 7)에서 종속변수인 체지방률과 상관계수 절대값이 0.5보다 큰 독립변수들은 신장, 윗몸일으키기, 악력(D), 악력(ND), 제자리멀리뛰기, 20m왕복오래달리기, 10m왕복달리기 7개로 관찰되었다. 머신러닝 단순회귀분

석에서 종속변수-체지방률, 독립변수-10m왕복달리기기록 1개로 정하여, 10m왕복달리기기록이 체지방률에 미치는 영향을 분석하였다. (Fig. 8)은 모집단에서 유효한 양적 데이터 변수(종속-독립) 간의 단순회귀모델 예측 분석 결과 사례를 나타낸다.

(Fig. 8)과 같이, 단순회귀모델 식이 체지방률(Y) = -0.0489 + 2.039*(10m 왕복달리기 기록)으로 도출되었다. 결정계수값이 0.3236으로 나와, 최대 1의 값과 비교했을 때 크지 않게 나와서 예측결과가 적합한 것은 아닌 것으로 판단되었다. 그러나, P<a (유의수준 0.05) 이므로 대립가설이 채택되었다. 이에 따라, 모집단의 회귀계수는 0이 아니다. 따라서 이 '10m 왕복달리기 ' 변수는 모집단에서 유효하다. 이에 따라, 본 단순회귀모델 식은 모집단에서 유효한 결과로 활용가능하다.

6.4 양적 데이터 변수 간 모집단에서의 상관성 여부와 다중 회귀 모델 예측 분석 사례

(Fig. 7)에서 종속변수인 체지방률과 상관계수 절대값이 0.5보다 큰 독립변수들은 신장, 윗몸일으키기, 악력(D), 악력(ND), 제자리멀리뛰기, 20m왕복오래달리기, 10m왕복달리기 7개로 관찰되었다. 머신러닝 다중회귀모델 분석을 위해, 10M달리기 변수와 상관성이 큰 윗몸일으키기와 제자리 멀리뛰기 변수를 제외하고, 악력D와 상관성이 큰 악력(ND) 변수까지 제외하였다. 이에 따라, 머신러닝 다중회귀모델 분석에서 종속변수-체지방률, 독립변수-신장, 악력(D), 20m 왕복오래달리기, 10m왕복달리기기록 4개로 정하여, 이 4개 독립변수들이 체지방률에 미치는 영향을 분석하였다.

회귀분석 통계량								
다중 상관계수	0.669314							
결정계수	0.447982							
조정된 결정계수	0.447467							
표준 오차	5.689081							
관측수	4292							
분산 분석								
	자유도	제곱합	제곱 평균	F 비	유의한 F			
회귀	4	112601.6	28150.41	869.7620097	0			
잔차	4287	138751.5	32.36565					
계	4291	251353.2						
	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	53.20239	2.497759	21.30005	8.88429E-96	48.30549	58.09929	48.305488	58.09929
신장	-0.18374	0.014835	-12.3856	1.21656E-34	-0.21282	-0.15465	-0.2128203	-0.15465
악력(D)	-0.06021	0.013159	-4.57572	4.88005E-06	-0.08601	-0.03441	-0.0860075	-0.03441
20m 왕복오래달	-0.13352	0.006914	-19.3099	9.87883E-80	-0.14707	-0.11996	-0.1470717	-0.11996
10m 왕복달리기	0.727613	0.05976	12.17567	1.49138E-33	0.610453	0.844772	0.6104531	0.844772

(Fig. 9) Cases of multiple regression model prediction analysis results between valid quantitative data variables (dependent-independent) in the population

(Fig. 9)는 모집단에서 유효한 양적 데이터 변수(종속-독립) 간의 다중회귀모델 예측 분석 결과 사례를 나타낸다. (Fig. 9)와 같이, 다중회귀모델 식이 체지방률(Y) = 53.20-0.1837*(신장)-0.06*(악력(D))-0.134*(20m왕복오래달리기)+ 0.728*(10m왕복달리기 기록)으로 도출되었다. 결정계수값이 0.448로 나와, 최대 1의 값과 비교했을 때 크지 않게 나와서 예측결과가 적합한 것은 아닌 것으로 판단되었다. 그러나, (Fig. 8) 단순회귀모델 식에서의 결정계수 0.3236보다는 향상된 결정계수를 나타내어, 본 다중회귀모델을 통해 예측결과의 적합도는 향상된 것으로 나왔다. 그리고 모든 독립변수들에 대해, P<a (유의수준 0.05) 이므로 대립가설이 채택되었다. 이에 따라, 모집단의 회귀계수는 0이 아니다. 따라서 이 4개 변수들은 모두 모집단에서 유효하다. 이에 따라, 본 다중회귀모델 식은 모집단에서 유효한 결과로 활용가능하다.

6.5 모집단에서 유효한 단일/다중 회귀예측모델을 의사결정에 활용한 사례

한국 성인인구 약 4천만명에 대한 체지방률과 10m왕복달리기 간의 단순 회귀 모델 식을 공공데이터 4292개 표본으로부터 도출하였다. 이 예측 식을 이용하여, 체지방률을 낮추기 위해 국민들에게 야외 운동을 장려하고, 기록 종목에 대한 관심과 참여를 확대할 수 있다. 또한, 체육시설의 확대에 있어, 10m 달리기 트랙 및 시간 기록 장비 관련시설 확충 정책을 펼칠 수 있다. 또한, 한국 성인인구 약 4천만명에 대한 체지방률과 3개 운동 종목 기록 간의 다중 회귀 모델 식을 공공데이터 4292개 표본으로부터 도출하였다. 이 예측 식을 이용하여, 국민들에게 자신의 신장을 고려하여, 체지방률을 낮추기 위해 자주 쓰는 손의 힘, 악력(D)와 야외 달리기 운동을 장려하고, 기록 종목에 대한 관심과 참여를 확대할 수 있다. 또한, 체육시설의 확대에 있어, 악력과 10m 및 20m 달리기 트랙 및 시간 기록 장비 관련시설 확충 정책을 펼칠 수 있다.

7. 본 기초 데이터과학 교육 사례의 효과성

<Table 3>과 <Table 4>는 스프레드시트를 활용한 기초 데이터과학 커리큘럼의 교육 만족도 결과를 나타

낸다. 34시간 교육과정에 대한 만족도 조사 영역으로 6개 항목을 선정하였다. 선정한 만족도 영역 항목은 데이터 특성 이해 교육, 공공데이터 수집을 통한 모집단의 평균값 예측 및 검정, 공공데이터 수집을 통한 모집단 그룹간 평균값차이 검정, 공공데이터 수집을 통한 모집단 내 질적 데이터의 연관성 예측 및 검정, 공공데이터 수집을 통한 모집단 내 양적 데이터의 상관성 예측 및 검정 그리고 공공데이터 수집을 통한 모집단 단일/다중 회귀 모델 검정과 같다.

<Table 3>은 20년도 2학기 30명 예비교사 대상 교육 만족도 조사 결과이며, <Table 4>는 동일 학기 30명 현직교사 대상 교육 만족도 조사 결과를 나타낸다. 교육통계학 과목을 학부 때 이수했던 현직교사들은 본 교육과정에 대해 70%이상 긍정적인 답변을 하였고, 예비교사 학부생들은 50%이상 긍정적인 답변을 하였다. 통계분석의 필요성을 숙지하고 있는 현직 교사들이 본 논문에서 제안한 기초 데이터과학 34시간 커리큘럼의 효과성을 보다 높게 평가하였다.

<Table 3> Satisfaction survey result of basic data science education for preservice teachers (30 person)

Satisfaction evaluation area (Training through public data collection)	very bad	bad	nor mal	good	very good
Training to understand data characteristics	0% (0)	10% (3)	20% (6)	40% (12)	30% (9)
Predicting and testing the average value of the population	0% (0)	10% (3)	30% (9)	30% (9)	30% (9)
Test of difference in mean value between population groups	0% (0)	10% (3)	30% (9)	30% (9)	30% (9)
Predicting and testing the association of qualitative data within a population	0% (0)	10% (3)	40% (12)	30% (9)	20% (6)
Predicting and testing the correlation of quantitative data within a population	0% (0)	10% (3)	20% (6)	30% (9)	40% (12)
Population single/multiple regression model test	0% (0)	10% (3)	20% (6)	40% (12)	30% (9)

<Table 4> Satisfaction survey result of basic data science education for field teachers (30 person)

Satisfaction evaluation area (Training through public data collection)	very bad	bad	nor mal	good	very good
Training to understand data characteristics	0% (0)	10% (3)	10% (3)	30% (9)	50% (15)
Predicting and testing the average value of the population	0% (0)	10% (3)	20% (6)	20% (6)	50% (15)
Test of difference in mean value between population groups	0% (0)	10% (3)	20% (6)	20% (6)	50% (15)
Predicting and testing the association of qualitative data within a population	0% (0)	10% (3)	20% (6)	30% (9)	40% (12)
Predicting and testing the correlation of quantitative data within a population	0% (0)	0% (3)	10% (3)	30% (9)	60% (18)
Population single/multiple regression model test	0% (0)	0% (3)	20% (6)	50% (15)	30% (9)

8. 결론

파이썬을 사용한 머신러닝 데이터과학 고급 교육과정에서는 학습자가 수집한 데이터를 저장 및 가공하는 과정을 시각적으로 한눈에 쉽게 이해하기 어렵다는 단점이 있다. 이를 보완하기 위해, 본 논문에서는 현장 교사 및 예비교사를 위한 데이터과학 소양교육으로 적용할 수 있는 기초 데이터과학 실습 교육 사례를 제시하였다. 그리고, 스프레드시트 데이터 분석 도구를 활용한 34시간 17주 교육 과정을 제안하였다. 데이터 수집, 데이터 가공 및 데이터 분석을 위한 도구로서, 스프레드시트는 파이썬과 달리, 데이터 분석 과정에서 있어서, 프로그래밍 및 자료구조 설정 과정이 복잡하지 않고 시각적으로 그 과정을 설명하고 이해할 수 있다. 물론, 머신러닝의 다양한 분석 알고리즘을 실행할 수 없는 단점이 있으나, 질적 데이터와 양적데이터에 대한 분석 이론을 습득하는 데 집중할 수 있는 장점이 있다.

본 논문에서 제안한 공공 빅데이터와 스프레드시트 활용 기초 데이터과학 교육 사례는 질적/양적 데이터에 대한 이해와 통계학의 필요성에 대한 이해를 높이는 데 초점을 두었다. 반면에 프로그래밍 활동을 줄인 것이다. 본 기초 데이터과학 교육을 이수한 후, 파이썬을 사용한 머신러닝 데이터과학 고급 교육과정을 이수하는 것이

타당하다. 제안한 기초 데이터과학 실습 커리큘럼은 현장교사와 예비교사들에게 데이터과학 입문단계로 적합하다. 프로그래밍 언어에 대한 부담이 없는 데이터분석 도구로 스프레드시트를 사용하여, 데이터 저장(수집), 가공 및 분석하는 절차에만 집중하여 교육할 수 있다.

참고문헌

[1] Ministry of Science and ICT(2017), The 4th Industrial Revolution in History, *R&D KIOSK*, 40.

[2] Ministry of Science and ICT(2017), The Various Aspects of the Fourth Industrial Revolution, the Realized Future, *R&D KIOSK*, 41.

[3] Kim, J.Y.(2016). *Hello Data Science*, Seoul : Hanbit Media.

[4] Kwon, J.K.(2020). *Learning data science*, Seoul : Jpub.

[5] Ichiro, U., Hiroaki, N., Masami, A. and Eichi, M.(2020). *Learning Data science with Excel*, Seoul : Hanbit Media.

[6] Seoul Open Data Plaza, 2021, [Online]. Available: http://data.seoul.go.kr.

[7] Public Data Portal, 2021, [Online]. Available: https://www.data.go.kr.

[8] National Statistics Portal, 2021, [Online]. Available: http://kosis.kr.

[9] Microdata integration service, 2021, [Online]. Available: http://mdis.kostat.go.kr.

[10] Jang, Y.J.(2017). Searching for the direction of data science education in the era of the 4th industrial revolution. *Integrated Humanities Research, 9*(10), 155-180.

[11] Park, Y.S. and Lee, S.J.(2020). Study on the Direction of Universal Big Data and Big Data Education-Based on the Survey of Big Data Experts, *Journal of The Korean Association of Information Education, 24*(2), 201-214.

[12] Hur, K.(2020), A Study on Elementary Education Examples for Data Science using Entry, *Journal of The Korean Association of Information Education, 24*(5), 473-481.

[13] Hong, J.Y. and Kim, Y.S.(2020), Development of AI Data Science Education Program to Foster Data Literacy of Elementary School Student, *Journal of The Korean Association of Information Education, 24*(6), 633-641.

[14] Kim, B.C. and et.al(2021), The Effect of Data Science Education on Elementary School Students' Computational Thinking: Focusing on Micro:bit's Sensor Function, *Journal of The Korean Association of Information Education, 25*(2), 337-346.

[15] Koo, D.H. and et.al(2020), Development of Data Science Education Program for Graduate School of AI Education, *The Korean Association of Information Education Research Journal, 11*(3), 15-26.

저자소개

허 경

1998년 고려대 전자공학과 학사

2000년 고려대 전자공학과 석사

2004년 8월 고려대 전자공학과 통신공학박사

2004년 8월 ~ 2005년8월 삼성종합기술원(SAIT) 전문연구원

2005년 9월 ~ 현재 경인교대 컴퓨터교육과 교수

관심분야: 피지컬컴퓨팅 SW교육, AI교육, 데이터과학교육

e-mail: khur@ginue.ac.kr