

## AI 활용도 200% 향상 #13

### AI로 데이터 분석하기 - part 1

이봉우

#### ■ AI로 데이터 분석하기란?

교육 데이터(지식·인식·행동, 배경변인 등)를 AI가 읽고 전처리·탐색·시각화·검정까지 도와주는 일련의 작업 흐름입니다.

- 입력: 엑셀·CSV·로그 파일, 간단한 설명(변수 뜻, 척도)
- 처리: 결측·오차 정리, 척도 수치화, 요약통계/그래프 생성, 가설검정(t, z, 회귀 등)
- 출력: 표·그림·요약문·정책용 원고 초안(APA 스타일 인용 등)

#### ■ AI로 분석을 하면 어떤 점에서 좋을까?

##### 1. 시간 절약과 효율성

- 반복 작업(전처리, 그래프, 표 서식)을 자동화해 핵심 판단에 집중할 수 있습니다.
- 논문·자료도 일관된 형식으로 요약해 탐색 시간을 단축합니다.

##### 2. 일관성·품질 관리

- 맞춤 지침(Custom Instructions)에 “학술 톤, APA 인용, 표/목록 중심 요약” 등을 저장하면, 매번 같은 품질로 결과물을 얻습니다.

##### 3. 재현성과 문서화

- 분석 단계(데이터 설명-처리-검정-해석)를 템플릿으로 고정하여, 수업·연구·행정 문서에 즉시 전환 가능합니다.

##### 4. 정책·수업 연결 용이

- 자동 생성되는 핵심 지표·그림을 바로 보고서/회의자료에 삽입하고, 수업 설계·PD·학교 운영 개선으로 신속히 연계할 수 있습니다.

## ■ 시작하기

데이터 분석은 프로그램(예: SPSS)을 이용하여 많이 하는 과정입니다. 이번에는 인공지능을 이용하여 데이터 분석을 하는 과정을 예로 설명해보려고 합니다.

우선 샘플 데이터가 필요하니, 제가 예전에 수집했던 데이터(에너지 소양-지식, 인식, 태도 등)의 일부를 사용해보겠습니다. (AI사용팁\_12\_data.xlsx)

이 데이터를 열어보면, 학년, 성별, 지식문항 10개에 대한 응답 결과, 태도, 행동 등에 대한 응답 결과들이 제시되어 있습니다.

[illegible]

## ■ 데이터에 대한 설명, 전처리

- 데이터를 업로드하면 AI가 알아서 데이터에 대한 해석을 합니다만, 어떤 데이터인지 설명을 해주면 좋습니다.
- 데이터 업로드와 함께 처음 해야 할일은 전처리입니다. 전처리를 하는 과정은 이미 연구를 진행하시면 많이 해보았을텐데, AI를 이용하면 좀 더 쉽게 할 수 있습니다. 물론 아직 완전하게 하지 못하는 경우가 있습니다. 이때에는 중간에 AI가 확인을 하도록 단계적으로 수행하게 하면 좋은 결과를 얻을 수 있습니다.
- 이번 예시에서는 아래의 내용과 같이 지식 문항에 대한 정답여부 확인, 리커트형 응답은 숫자로 값 변환 등을 수행합니다. 이 데이터에는 결측값이 없는데, 결측값이 있는 경우에는 어떻게 해야 할지 제시해줄 필요가 있습니다.

이 데이터는 고등학생의 에너지 소양을 측정한 결과를 나타낸 데이터입니다. 학년, 성별이 표시되어 있고, 에너지 소양 관련 질문에 대한 학생들의 응답이 제시되어 있습니다. 에너지 소양 관련 질문은 에너지 지식 문항 10개, 에너지에 대한 인식 5개, 에너지 관련 행동 5개로 구성되어 있습니다.

우선 이 데이터를 전처리해줘.

1. 에너지에 대한 지식 문항에 대한 정답 여부를 담은 열을 추가해 주세요.

에너지에 대한 지식의 정답은 다음과 같습니다.

- 1 - 일을 할 수 있는 능력
- 2 - 태양 에너지
- 3 - 킬로와트시 (kWh)
- 4 - 전력 X 사용한 시간
- 5 - 100원
- 6 - 화학 에너지 → 전기 에너지 → 빛 에너지
- 7 - 열에너지
- 8 - 화학 에너지
- 9 - 같다.
- 10 - 100의 에너지를 투입하면 30의 전기 에너지가 생산된다.

2. 에너지에 대한 인식, 행동 질문은 리커트 척도형으로 되어 있습니다. 응답을 숫자로 바꿔주세요. 매우 그렇다(5), 그렇다(4), 보통이다(3), 아니다(2), 매우 아니다(1)로 바꾸어주세요.

전처리의 전 과정을 한국어로 설명해주고, 이 전처리된 결과물을 확인할 수 있게 다운로드 링크로 제공해 주세요.

- 결과의 일부만 제시해보았습니다. 전처리를 잘 했습니다.

전처리 과정을 완료했습니다. 아래는 처리 과정의 단계별 설명입니다.

◆ 1. 데이터 구성

원본 데이터에는

- 학년, 성별,
- 에너지 지식 문항(`concept01~10`),
- 에너지 인식 문항(`interest01~05`),
- 에너지 행동 문항(`behavior01~05`)

이 포함되어 있었습니다.

◆ 2. 에너지 지식 문항 정답 처리

각 지식 문항별로 제시된 정답 기준과 학생 응답을 비교하여,  
정답이면 1, 오답이면 0으로 표시한 새로운 열( `concept01_정답여부` 등)을 추가했습니다.

예시:

문항	정답	응답	정답여부
concept01	일을 할 수 있는 능력	일을 할 수 있는 능력	1
concept02	태양 에너지	지구 내부의 열에너지	0

◆ 3. 인식 및 행동 문항 리커트 척도 변환

리커트 척도형 문항(예: “매우 그렇다”, “그렇다” 등)을 다음의 숫자 값으로 변환했습니다.

원문	변환값
매우 그렇다	5
그렇다	4
보통이다	3
아니다	2
매우 아니다	1

■ 탐색적 데이터 분석(EDA, exploratory data analysis) : 일변량/다변량

- 처음 출발은 기본 통계값을 분석해달라고 합니다. 이는 일변량 탐색적 데이터 분석을 통해 할 수 있습니다.
- 요즘에는 데이터만 입력해도 자동으로 다음 과정에 대한 제안을 하니까, AI의 제안대로 따라해도 됩니다.

이 데이터에 대해서 깊이 있게 이해하기 위해 EDA를 실시하고자 합니다. 우선 일변량 EDA부터 실시해 주세요.

## 1) 분석 개요

- **대상 데이터:** 전처리(리커트 수치화 및 지식 정답여부 추가)된 `energy_literacy_processed_v2.xlsx` 를 기반으로 재계산하여 EDA 수행
- **변수 유형**
  - 범주형: 학년, 성별
  - 지식 정답여부: `concept01_정답여부` ~ `concept10_정답여부` (0/1)
  - 리커트(숫자화): `interest01` ~ `interest05`, `behavior01` ~ `behavior05` (1~5)
  - 파생점수: `knowledge_score_sum` (0~10 합계), `interest_score_mean` (1~5 평균), `behavior_score_mean` (1~5 평균)

## 2) 제공 산출물(다운로드)

-  **단변량 EDA 리포트(엑셀, 다중 시트):**  
`energy_univariate_report.xlsx`
  - `info`: 데이터 크기
  - `missingness`: 컬럼별 결측치 개수/비율
  - `categorical_dist`: 학년/성별 분포
  - `likert_summary`: 리커트 항목별 요약통계(N, 평균, 표준편차, 분위수 등)
  - `likert_value_counts`: 리커트 값(1~5)과 NaN 분포
  - `correctness_summary`: 지식 정답여부(0/1) 요약통계
  - `correctness_value_counts`: 정답/오답/NaN 분포
  - `derived_scores_summary`: 파생 점수 3종 요약통계
-  **주요 시각화(이미지):**
  - 학년 분포(막대)
  - 성별 분포(막대)
  - 지식 점수 합 분포(히스토그램)
  - 인식 평균 분포(히스토그램)
  - 행동 평균 분포(히스토그램)

주의: 일부 그림에 한글 폰트 경고가 있을 수 있으나, 파일 자체는 정상 생성되었습니다.

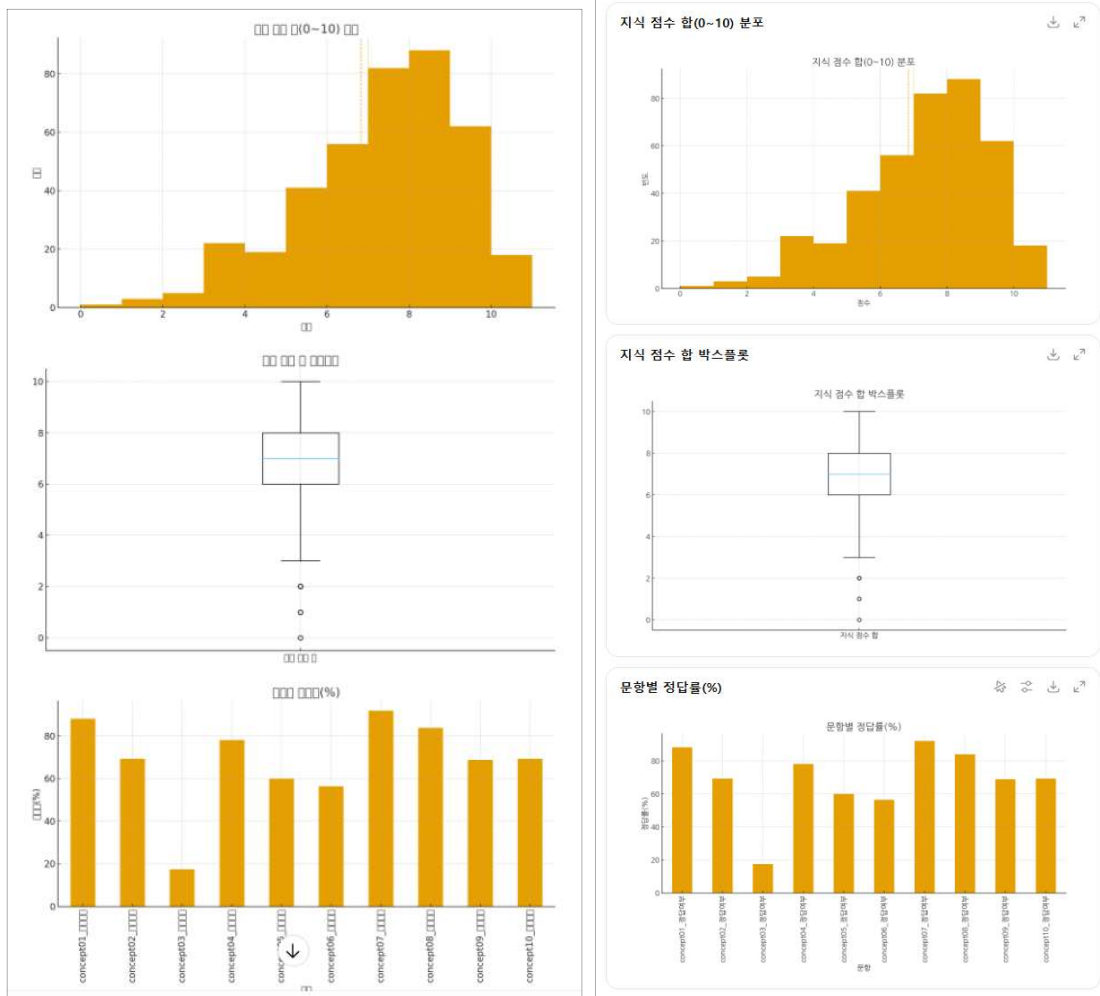
## ■ 시각화

- 시각화도 AI의 큰 장점입니다. 파이썬을 이용하여 분석하기 때문에 파이썬에 있는 시각화 라이브러리를 사용할 수 있습니다. seaborn의 시각화를 참고하여 어떤 것이 가능한지 살펴보면 좋을 것 같습니다. (<https://seaborn.pydata.org/> 참조)
- 아쉽게도 바로 입력하면 그래프에 한글이 나오지 않습니다. 이를 해결하는 방법이 여러 가지 있지만 저는 다음과 같은 방법을 사용합니다.

(koreanize\_matplotlib-0.1.1-py3-none-any.whl)

# 지식 점수의 결과를 시각적으로 보여 주세요.

# Code interpreter를 활용해서 내가 제공한 whl file을 다운로드하여, 너의 가상환경에 설치해줘. 그런 다음에 한국어 사용 설정을 해줘. 한국어가 정상적으로 표시되도록 한 후 위 차트를 다시 그려 줘.



# 다변량 EDA를 실시해 주세요.

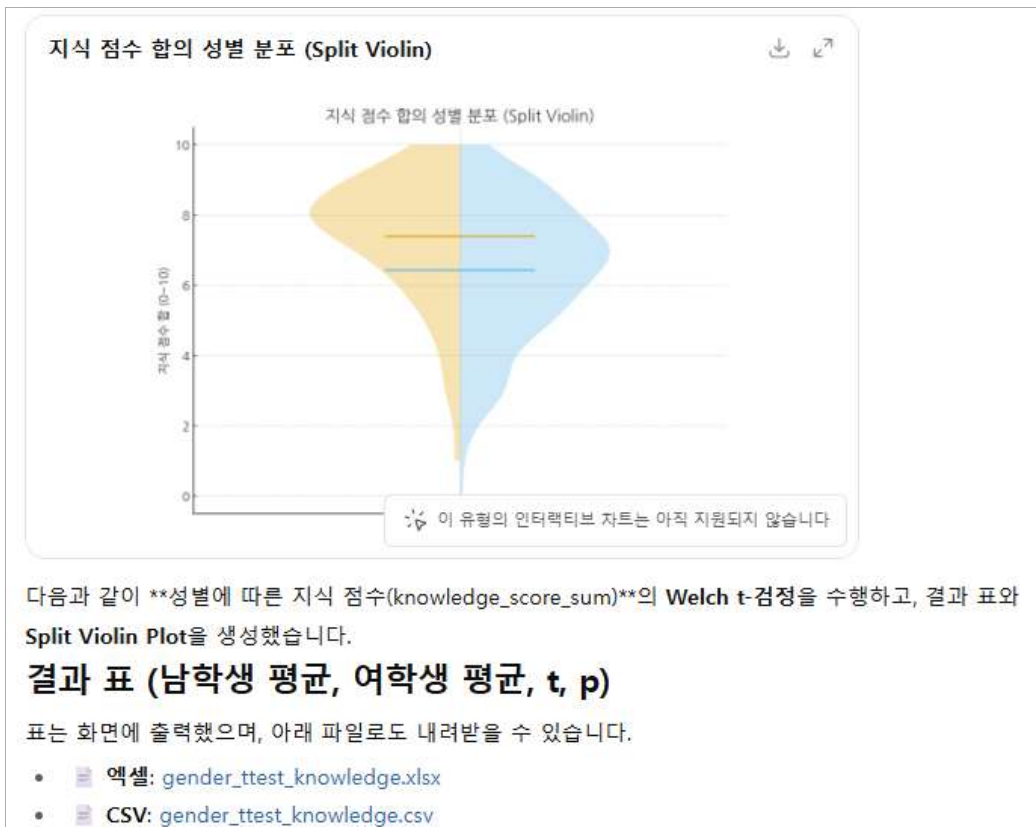
- 여러 결과를 보여줍니다. (생략)

## ■ 분석

- 데이터 업로드하고, EDA를 실시하는 과정에서 사실 간단한 분석은 이미 AI가 제안하기 때문에 그대로 수행해도 되지만, 특별히 원하는 분석, 원하는 시각화 결과들이 있으면 이를 제안하면 됩니다. 예시로 몇 개 해보겠습니다.

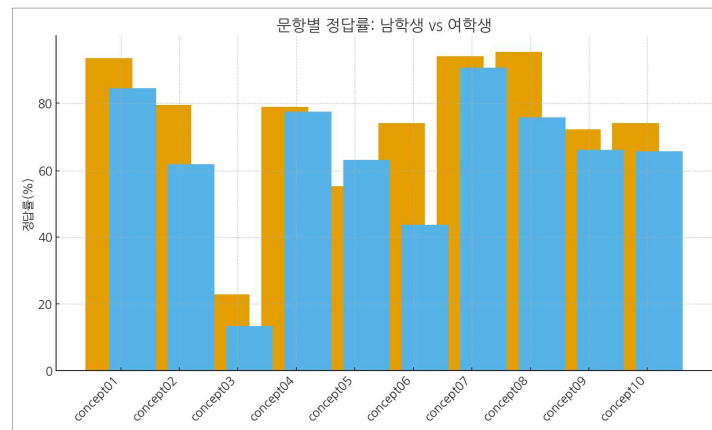
에너지에 대한 지식 결과를 성별에 따른 차이로 t-검정해주세요.

결과를 남학생 평균, 여학생 평균, t값, p-value와 같이 정리해서 표로 제시해주고, 다운받을 수 있게 해주세요. 위 결과를 바이올린플롯으로 시각화해 주세요.



## 문항별로 분석해줘.

- 결과를 문항별로 보고 싶어서 문항별로 분석해달라고 했더니, t-검정 대신에 z-검정이 타당하다는 제안과 함께 분석 결과를 보여줍니다.
- 문항별로 성별에 따른 정답률 차이를 분석했습니다. 이분형(정답/오답) 변수에 대해 t-검정 대신 두 집단 비율 차이 z-검정을 적용하는 것이 통계적으로 타당하여, 각 문항에 대해 남학생·여학생 정답률, 차이, z, p-value를 산출하고 \*\*다중비교 보정(FDR, Benjamini-Hochberg)\*\*까지 제공했습니다.



문항	정답률 (%)	z	p_value	q_value(FDR)
1	concept01	2.725	0.006428620985043709	0.016071552462609273
2	concept02	3.752	0.00017550697689383554	0.0005850232563127852
3	concept03	2.457	0.014029156606252702	0.028058313212505404
4	concept04	0.339	0.734678889491351	0.734678889491351

문항별로 성별에 따른 정답률 차이를 분석했습니다. 이분형(정답/오답) 변수에 대해 t-검정 대신 두 집단 비율 차이 z-검정을 적용하는 것이 통계적으로 타당하여, 각 문항에 대해 남학생·여학생 정답률, 차이, z, p-value를 산출하고 \*\*다중비교 보정(FDR, Benjamini-Hochberg)\*\*까지 제공했습니다.

### 제공물

- 표(엑셀): itemwise\_gender\_test\_knowledge.xlsx
- 표(CSV): itemwise\_gender\_test\_knowledge.csv
- 보조 시각화(문항별 정답률 막대): itemwise\_gender\_accuracy\_bar.png

표에는 다음 열이 포함됩니다.

- 문항(concept01~concept10)
- 남학생\_N, 여학생\_N
- 남학생\_정답률(%), 여학생\_정답률(%), 차이(남-여, %)
- z, p\_value, q\_value(FDR)



## ■ 과제 제안 요청

- 이 데이터를 이용하여 할 수 있는 과제(인사이트)를 제안할 수도 있습니다. 지금 예시로 한 데이터는 원래의 데이터에서 많이 줄여놓은 것이기에 할 수 있는 제안이 그리 많지 않지만, 그래도 놀라운 인사이트를 보여줍니다.

이 데이터를 기반으로 에너지 교육과 관련된 교육적 인사이트를 도출하고 싶어. 분석할 만한 과제를 제안해 줘.

## <결과 (일부)>

아래 제안들은 현재 보유하신 \*\*에너지 소양 데이터(지식·인식·행동, 학년·성별 포함)\*\*를 바탕으로, 교육적 인사이트를 체계적으로 도출하기 위한 분석 과제 포트폴리오입니다. 각 과제는 연구질문 → 가설 → 분석법 (통계 절차) → 시각화/산출물 → 교육적 해석 포인트로 구성했습니다. 필요하시면 항목별로 바로 실행 가능한 코드/워크플로로 세분화해 드리겠습니다.

### A. 측정 도구의 신뢰도·구조 타당도 검증

#### A-1. 신뢰도(내적 합치도) 및 문항 진단

- 연구질문: 인식·행동 척도가 일관되게 구성 개념을 측정하는가? 지식 문항의 난이도·변별도는 적절한가?
- 가설: 인식/행동 척도의 Cronbach's  $\alpha \geq .70$ ; 지식 문항은 난이도(정답률 .2~.8 범위), 변별도(상·하위 집단 간 차) 양호.
- 분석: Cronbach's  $\alpha$ , 문항-총점 상관, 지식 문항 정답률·상·하위 27% 그룹 차이.
- 시각화/산출물: 신뢰도 표, 문항-총점 상관 히스토그램, 지식 문항 난이도-변별도 산포도.
- 교육적 해석: 난이도 극단 문항(너무 쉽/어려움) 재구성, 변별도 낮은 문항의 선지 개선 및 수업 보완 포인트 도출.

### 참고문헌(예시, APA)

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2), 111–127.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances. *American Psychologist*, 50(9), 741–749.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

#### # Role

당신은 과학교육자, 에너지교육 전문가입니다. 지속가능한발전을 위한 에너지교육에 관심이 있고, 관련된 전문성이 있습니다.

#### # Objective

데이터에 근거해서 국가 차원의 에너지 교육 정책을 만들려고 합니다. 바로 실행 가능하고 효율적이며 효과적인 실행계획을 제안해 주세요.

<결과 (일부)>

## 1. 정책 목표(3개)와 핵심 원리(4개)

### 정책 목표

1. **핵심 개념** 숙달: 전력/에너지 구분, 단위(kWh), 효율·손실, 비용추정, 시스템적 사고.
2. **행동 전이** 촉진: 학교·가정에서의 절전·효율·수요관리·재생에너지 이해에 기반한 실천 행동 증대.
3. **지속가능 삶의 역량**: 데이터 해석, 의사결정, 시민참여(시민과학·지역 과제) 역량 강화.

### 핵심 원리

- 지식→인식(가치·흥미)→행동의 매개 경로를 설계(지식만 강화해선 행동 변화가 약함).
- 문항별 취약지점 정밀 보정(단위/전력·에너지 혼동, 효율 계산, 비용 산정 등).
- 증거 기반 적응형 지원(학교별·집단별 데이터로 맞춤 PD·콘텐츠 배포).
- 간편·저비용·즉시성: 교실 수업·학교 운영·지역 협력에서 바로 쓰는 도구 중심.

## 2. 실행 패키지(6종)

### A. 핵심 개념 모듈(주 1차시, 8-10차시 완결)

- **모듈 구성:**
  1. 에너지 보존·전환(열·화학·전기)
  2. 전력 vs 에너지, kWh(취약 문항 직접 타격)
  3. 전기요금·비용 추정(사용량→시간×전력, kWh×요금)
  4. 효율·손실("100 투입→30 산출" 유형)
  5. 수요관리·피크·부하곡선
  6. 재생에너지·혼합 포트폴리오
  7. 시스템 사고(가정-학교-지역-국가)
  8. 의사결정 시나리오(경제성/환경성/형평성 트레이드오프)
- **교수전략:** 오개념 정정(대비 사례·오답 토론), 미세 연습문항(단위 변환·요금 계산), 실험·모형 활동.