동국대학교 dongguk university

R E P O R T

EDA on Student Study Performance

• 과 목 명	: 인 공 지 능
• 담 당 교 수	: 박 소 현 교수님
• 제 출 일	: 2024-04-24
• 학 과	: 컴퓨터공학전공
• 학 번	: 2019212978
• 성 명	: 고 영 민

동국대학교 dongguk university

목 차

동국대학교 dongguk university

1. 개요

I. 선정 데이터 세트

- 학생들의 시험 성적 데이터

II. 선정 이유

- 해당 데이터 세트에는 성별, 인종 및 민족, 부모의 학력, 시험 전 점심 식사 여부, 시험 준비 과정, 수학 점수, 독해 점수 및 작문 점수가 포함되어 있다. 나는 교육학 및 사회학에 흥미가 있는데, 본 데이터를 통해 과목 간의 상관관계, 성적과 부모의 학력 간의 상관관계, 그리고 시험 직전의 식사가 시험 점수에 영향을 미치는지 등 다양한 변수 간의 상관관계를 탐색적 데이터 분석을 통해 밝혀 향후 교육 관련 프로젝트를 진행할 때 이를 반영하고 싶었기 때문에 본 데이터 세트를 선정하게 되었다.

III. 데이터 세트 출처

- 캐글(Kaggle)의 Student Study Performance 데이터 세트[1]

IV. 분석 도구

- Pandas, NumPy, Seaborn, Matplotlib

2. 데이터 세트 설명

2.1. 데이터 소개

전반적인 데이터 확인

```
2]: display(df)
```

	gender	race_ethnicity	parental_level_of_education	lunch	test_preparation_course	math_score	reading_score	writin
0	female	group B	bachelor's degree	standard	none	72	72	
1	female	group C	some college	standard	completed	69	90	
2	female	group B	master's degree	standard	none	90	95	
3	male	group A	associate's degree	free/reduced	none	47	57	
4	male	group C	some college	standard	none	76	78	
...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99	
996	male	group C	high school	free/reduced	none	62	55	
997	female	group C	high school	free/reduced	completed	59	71	
998	female	group D	some college	standard	completed	68	78	
999	female	group D	some college	free/reduced	none	77	86	

1000 rows × 8 columns

그림 1. 상위/하위 5 개 데이터 출력

- 학생의 성적(시험 점수)이 성별, 인종, 부모의 교육 수준, 점심 식사 및 시험 준비 과정과 같은 다른 변수에 의해 어떻게 영향을 받는지 이해하기 위해 학생들이 다양한 과목에서 획득한 점수로 구성되어 있다.
- 본 데이터 세트는 8 개의 속성과 1000 개의 데이터로 이루어져 있다.

데이터 요약표 확인

```
df.describe()
```

	math_score	reading_score	writing_score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

그림 2.각 과목별 점수 데이터의 데이터 요약표

- 3 개의 속성은 int64 타입이고, 5 개의 속성은 object 타입으로 구성되어 있다.
- 본 데이터 세트의 핵심 데이터인 수학 점수, 독해 점수, 작문 점수의 수치형 데이터 요약표는 그림 2 와 같다. 수학 점수의 경우, 평균은 66 점, 표준 편차는 15 점, 최솟값과 최댓값은 각각 0 점과 100 점이다. 독해 점수의 경우, 평균은 69 점, 표준 편차는 14 점, 최솟값과 최댓값은 각각 17 점과 100 점이다. 작문 점수의 경우, 평균은 68 점, 표준 편차는 15 점, 최솟값과 최댓값은 각각 10 점과 100 점이다.

2.2. 어노테이션 포맷

항목명	데이터 타입	설명
gender	object	학생의 성별
race_ethnicity	object	학생의 인종(그룹 A~E)
parental_level_of_education	object	부모의 교육 수준
lunch	object	시험 전 점심 식사 여부
test_preparation_course	object	시험 준비 정도
math_score	int64	수학 점수
reading_score	int64	독해 점수
writing_score	int64	작문 점수

I. **gender**
gender 속성은 학생의 성별이 포함된 열로 속성 값에는 'female', 'male'이 포함되어 있다. 데이터 타입은 object 이다.

II. **race_ethnicity**
race_ethnicity 속성은 학생의 인종이 포함된 열로 속성 값에는 'group A', 'group B', 'group C', 'group D', 'group E'가 포함되어 있다. 데이터 타입은 object 이다.

III. **parental_level_of_education**
partner_level_of_education 속성은 부모의 교육 수준(학력)이 포함된 열로 "bachelor's degree", 'some college', "master's degree", "associate's degree", 'high school', 'some high school'이 포함되어 있다. 데이터 타입은 object 이다.

IV. **lunch**
lunch 속성은 시험 전 점심 식사 여부가 포함된 열로 'standard', 'free/reduced'가 포함되어 있다. 데이터 타입은 object 이다.

V. **test_preparation_course**
test_preparation_course 속성은 시험 준비 정도가 포함된 열로 'none', 'completed'가 포함되어 있다. 데이터 타입은 object 이다.

VI. math_score & reading_score & writing_score
math_score, reading_score, writing 속성은 각각 수학, 독해, 작문 점수가 포함되어 있다. 데이터 타입은 int64 이다.

동국대학교 dongguk university

3. 데이터 전처리

3.1. 데이터 정제

결측치(missing values) 처리

```
df.isnull().sum()
```

```
gender                       0
race_ethnicity               0
parental_level_of_education  0
lunch                        0
test_preparation_course      0
math_score                   0
reading_score                0
writing_score                0
dtype: int64
```

- 이 데이터 세트에는 결측치가 존재하지 않는 것으로 나타났다.

중복값(duplicate values) 처리

```
df.duplicated().sum()
```

```
0
```

- 이 데이터 세트에는 중복값이 존재하지 않는 것으로 나타났다.

3.2. 데이터 변환

데이터 세트의 열 이름 확인

```
df.columns
```

```
Index(['gender', 'race_ethnicity', 'parental_level_of_education', 'lunch',
       'test_preparation_course', 'math_score', 'reading_score',
       'writing_score'],
      dtype='object')
```

데이터 세트의 열 이름 변경

Rename some columns

```
df = df.rename(columns = {'race_ethnicity' : 'group',
                          'parental_level_of_education' : 'parent_education_level',
                          'test_preparation_course' : 'test_preparation'})
df.columns
```

```
Index(['gender', 'group', 'parent_education_level', 'lunch',
       'test_preparation', 'math_score', 'reading_score', 'writing_score',
       'total_score', 'mean_score'],
      dtype='object')
```

- 식별성이 떨어지는 3 개의 열 이름을 각각 'group', 'parent_education_level', 'test_prepartation'으로 변경하였다.

파생변수(derived variable) 생성

Create derived variable

```
df['total_score'] = df['math_score'] + df['reading_score'] + df['writing_score']
df['mean_score'] = round(df['total_score'] / 3, 2)
df.head()
```

	gender	race_ethnicity	parental_level_of_education	lunch	test_preparation_course	math_score	reading_score	writing_score	total_score	mean_score
0	female	group B	bachelor's degree	standard	none	72	72	74	218	72.67
1	female	group C	some college	standard	completed	69	90	88	247	82.33
2	female	group B	master's degree	standard	none	90	95	93	278	92.67
3	male	group A	associate's degree	free/reduced	none	47	57	44	148	49.33
4	male	group C	some college	standard	none	76	78	75	229	76.33

- 여러 속성 간의 상관관계를 확인할 때 사용하기 위해 총합 점수와 평균 점수에 대해 파생변수를 생성하였다.
- 총합 점점은 수학 점수, 독해 점수, 작문 점수의 합이고, 평균 점수는 총합 점수를 3 으로 나누고 소수점 둘째자리에서 반올림 한다

데이터 결합

1. Check values on parent education level column

```
df['parental_level_of_education'].unique()
```

```
array(["bachelor's degree", 'some college', "master's degree",
       "associate's degree", 'high school', 'some high school'],
      dtype=object)
```

2. Convert the some high school to high school

Convert the some high school to high school

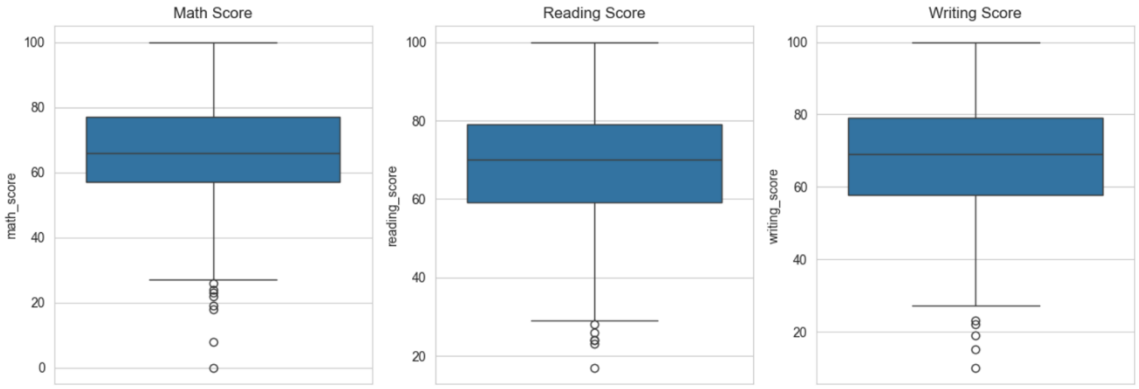
```
# some high school replace with high school
df['parent_education_level'] = df['parent_education_level'].replace('some high school', 'high school')
df['parent_education_level'].value_counts()
```

```
parent_education_level
high school           375
some college          226
associate's degree    222
bachelor's degree     118
master's degree        59
Name: count, dtype: int64
```

- 이 데이터 세트의 부모 교육 수준 속성에는 'high school' 속성값과 'some high school'이 존재하는데 고등학교라는 점은 동일하기 때문에 'some high school' 값을 replace() 함수를 통해 'high school'로 대체하였다.

이상치 제거

1. Check the distributions of scores by subject



2. Drop outliers

```
# Drop outliers
df.drop(df[df['math_score'] <= 20].index, inplace = True)
df.drop(df[df['reading_score'] <= 20].index, inplace = True)
df.drop(df[df['writing_score'] <= 20].index, inplace = True)
```

- 과목별 상자그림을 확인한 결과, 과목별 공통적으로 20 점 이하 값에서 이상치가 판별되었다. 따라서 20 점 이하의 값은 outlier 로 판단하고 분석하고자 하는 데이터에서 제거했다.

3.3. 데이터 정규화

데이터 정규화

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

columns_to_normalize = ['math_score', 'reading_score', 'writing_score', 'total_score', 'mean_score']
df[columns_to_normalize] = scaler.fit_transform(df[columns_to_normalize])
```

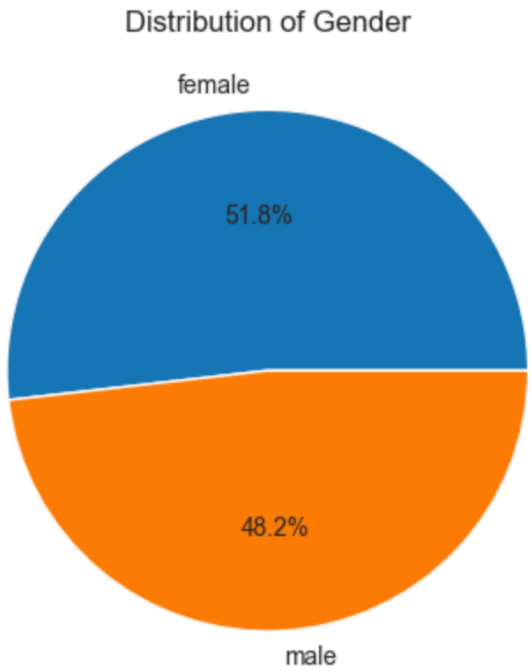
```
df.describe()
```

	math_score	reading_score	writing_score	total_score	mean_score
count	9.940000e+02	9.940000e+02	9.940000e+02	9.940000e+02	9.940000e+02
mean	4.628535e-16	-5.718654e-17	3.466934e-16	-7.505733e-17	7.398508e-16
std	1.000503e+00	1.000503e+00	1.000503e+00	1.000503e+00	1.000503e+00
min	-3.020862e+00	-3.049398e+00	-3.135590e+00	-3.042101e+00	-3.042101e+00
25%	-6.387098e-01	-6.619852e-01	-6.994983e-01	-7.029302e-01	-7.031667e-01
50%	-2.615643e-02	4.019524e-02	4.486320e-02	4.463973e-02	4.488704e-02
75%	7.225199e-01	7.248211e-01	7.215554e-01	7.138354e-01	7.139032e-01
max	2.287934e+00	2.146736e+00	2.142609e+00	2.311465e+00	2.311475e+00

- 평균=0, 표준편차=1 로 조정해서 모든 특성이 같은 크기를 갖도록 Standard Scaler 를 사용하여 'math_score', 'reading_score', 'writing_score', 'total_score', 'mean_score' 열의 데이터 값을 정규화하였다.

4. 데이터 시각화

4.1. 해당 데이터셋의 남녀 비율 확인

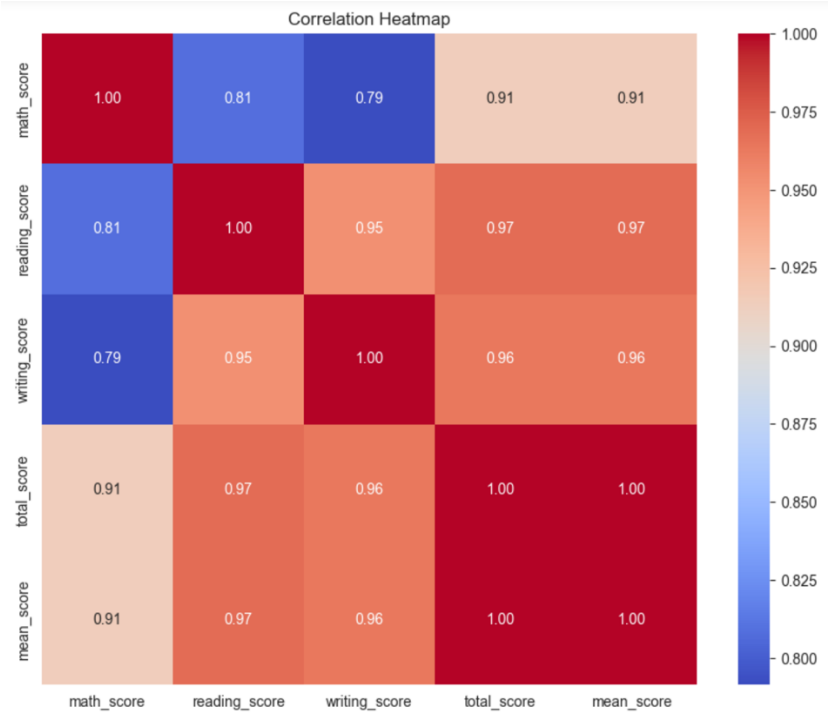


파이 차트는 전체 대비 각 부분의 비율을 원형 그래프로 나타낸 시각화 기법이다.

본 데이터셋을 분석할 때 성별 편향이 발생하는지 확인하기 위해 남성과 여성의 비율을 파이 차트를 사용하여 시각화 하였다.

시각화 결과, 남성이 48.2%를 차지하고 여성이 51.8%를 차지함을 확인했다. 따라서 데이터 분석 시 성별 편향이 분석에 영향을 미치지 않았음이 확인되었다.

4.2. 수치형 데이터들의 상관관계 확인

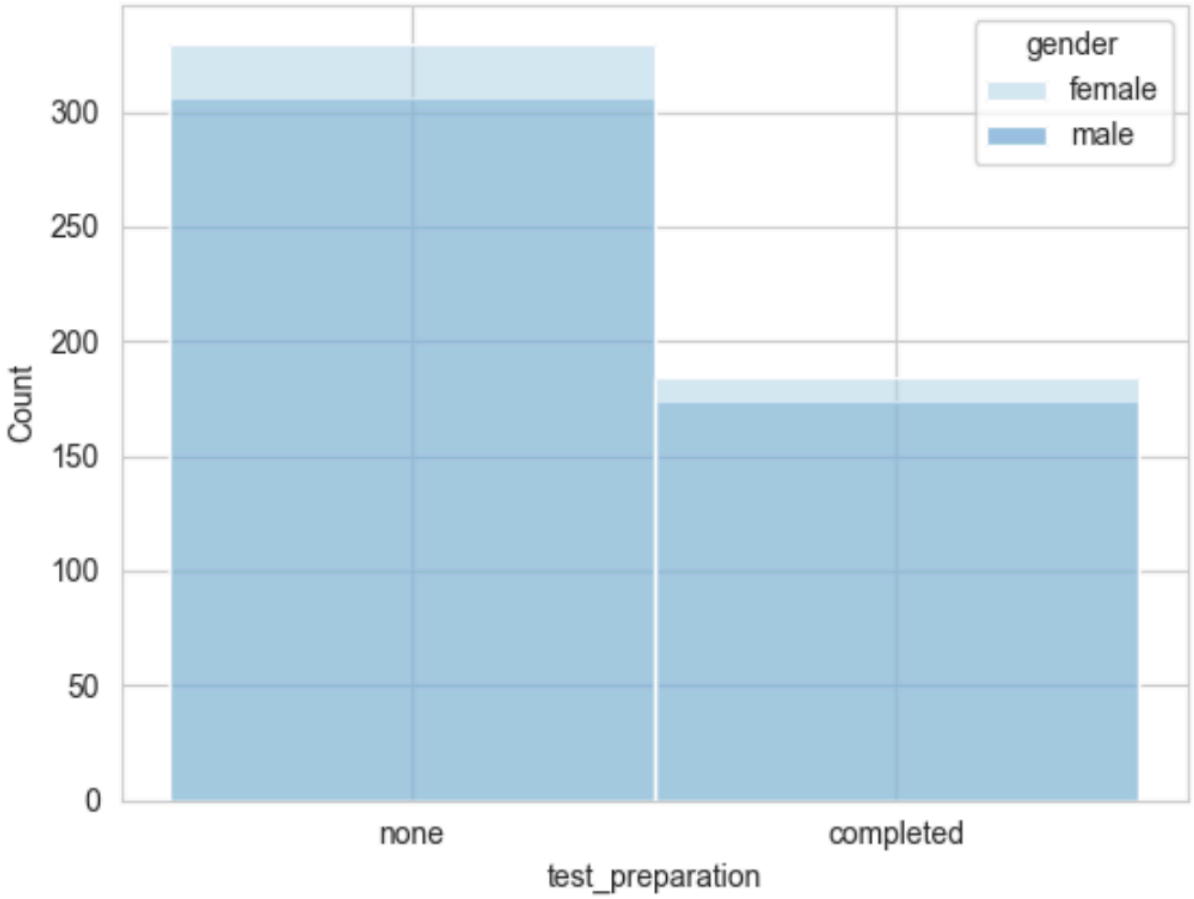


히트맵은 데이터의 상관관계를 색상으로 표현하는 시각화 기법이다. 본 시각화에서 사용한 히트맵은 상관관계가 높을수록 붉은색에 가까워지고, 상관관계가 낮을수록 파란색에 가까워진다.

수치형 데이터인 수학 점수, 독해 점수, 읽기 점수, 총점, 평균 점수에 대해 히트맵으로 시각화 하여 각 속성별 상관관계를 확인했다.

시각화한 히트맵의 결과를 분석하면, 최소 0.79 이상의 상관관계를 보여주기 때문에 모든 과목 간의 상관관계가 높다고 볼 수 있으며, 총점과 평균점수와 상관관계가 가장 높은 것은 독해 및 작문 점수였다. 또한 수학 점수는 독해 및 작문 점수와 높은 상관관계를 보였다. 그러나 가장 높은 상관관계를 보인 것은 독해와 작문 간의 상관관계이다.

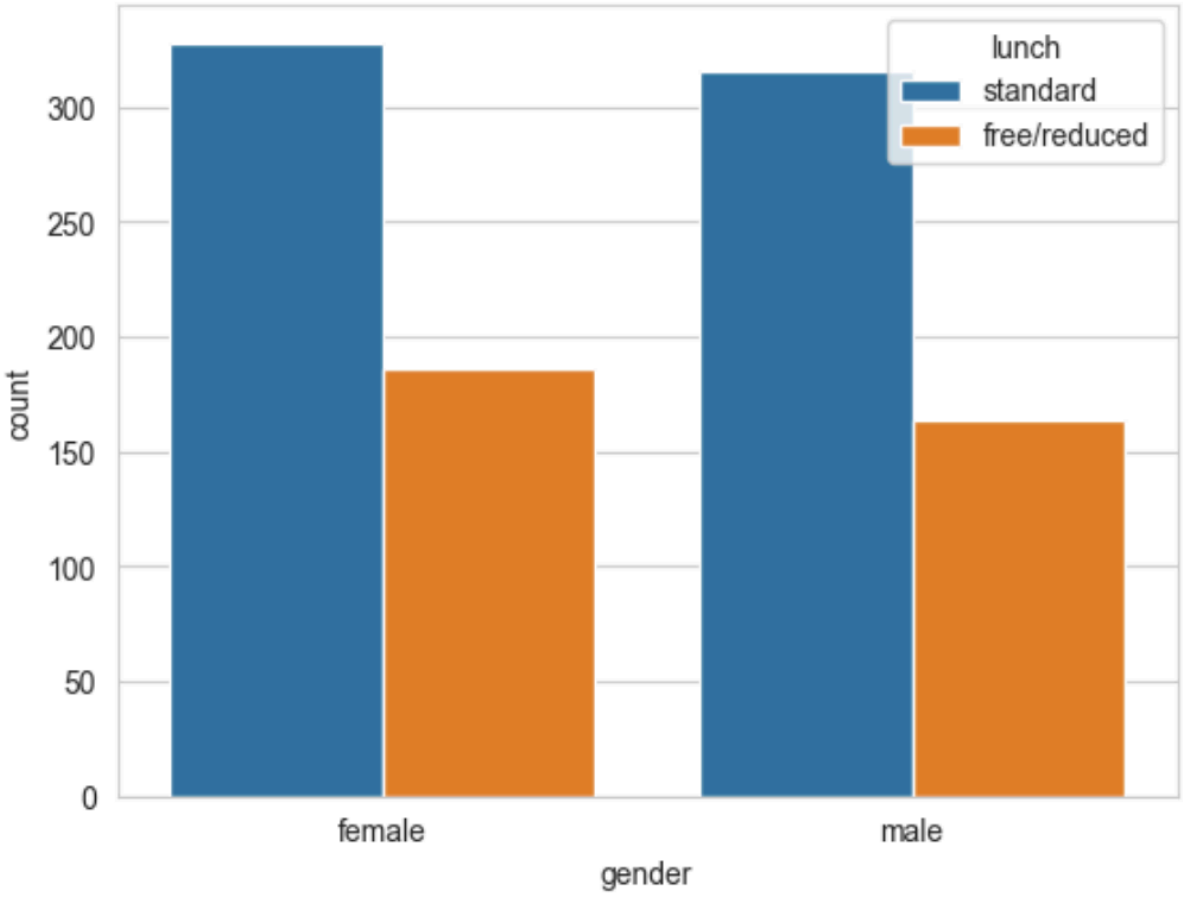
4.3. 성별에 따른 시험 준비 정도 확인



test_preparation 속성에 대해 여자의 경우 남자보다 더 높은 비율로 시험 준비가 되었을 것이라고 가정하고 성별에 따른 시험 준비 정도를 카운트 플롯으로 시각화 하였다.

시각화 결과를 분석하면, 여자가 남자보다 약간 더 많이 준비를 끝냈다는 결과를 확인할 수 있었다. 하지만, 예상과는 다르게 여자는 남자보다 준비를 못 끝낸 비율도 더 높았다.

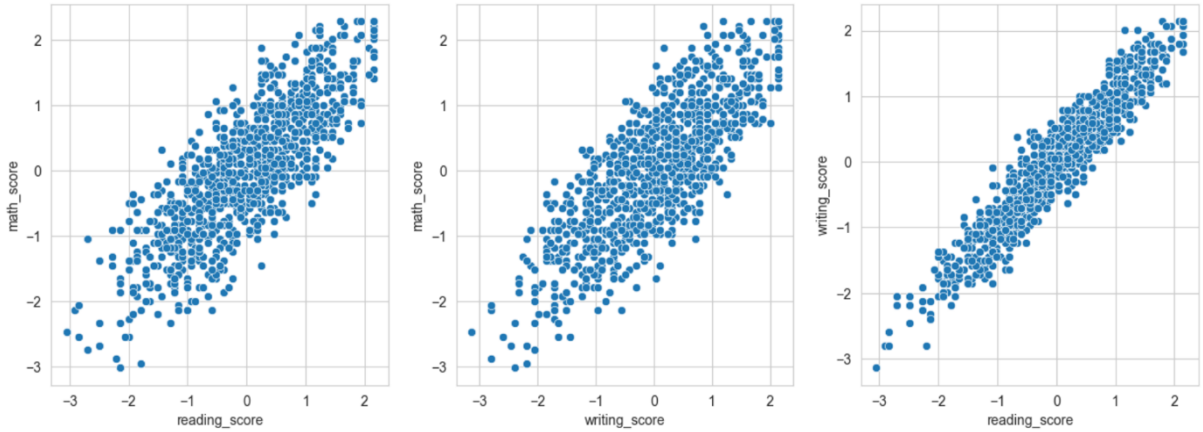
4.4. 성별에 따른 시험 전 식사 여부 확인



성별에 따른 시험 전 식사 여부를 확인하기 위하여 카운트 플롯을 활용하여 성별에 따른 식사 여부를 시각화 하였다. 이때 식사 여부에 따라 색상을 다르게 표시되도록 해 구분하였다.

시각화 결과를 분석하면, 남녀 집단 모두 일반식을 먹은 비율과 굶거나 적게 먹은 비율이 비슷했다. 즉, 남녀 집단 간 시험 전 식사 여부에 유의미한 차이가 없음을 나타낸다.

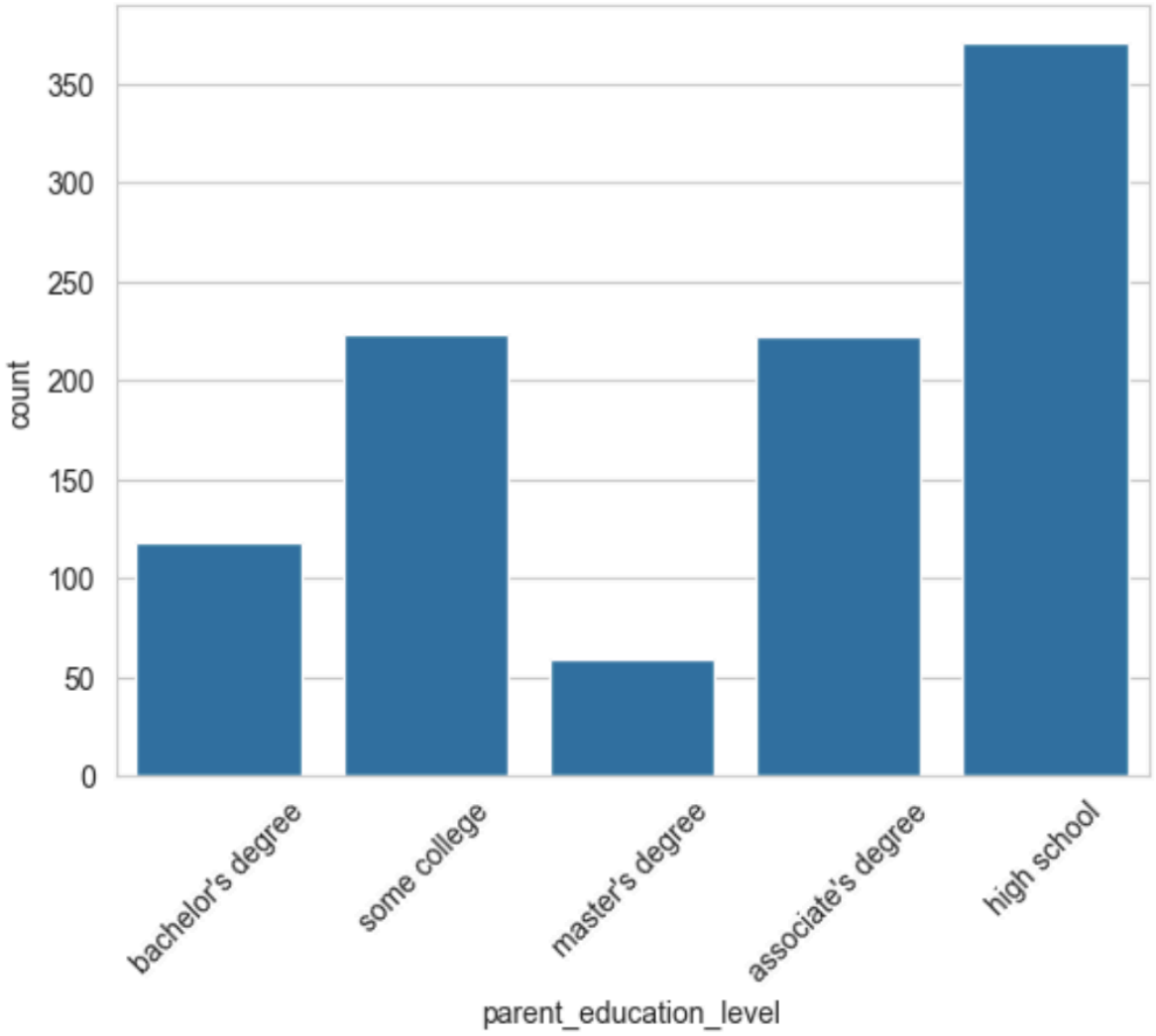
4.5. 각 과목별 산점도 그래프



4.2 에서 수치형 데이터들의 상관관계를 확인하였다. 4.5 에서는 각 과목 간의 상관관계를 명확히 파악하기 위해 각 과목별 산점도 그래프를 시각화 하였다.

먼저, 첫 번째는 독해 점수에 따른 수학 점수의 산점도 그래프이고, 두 번째는 작문 점수에 따른 수학 점수의 산점도 그래프이며, 마지막 그래프는 독해 점수에 따른 작문 점수의 산점도 그래프이다. 일반적으로 우상향의 분포는 높은 양의 상관관계가 있음을 나타내는데 3 개의 그래프에서 모두 높은 상관관계가 있음을 확인할 수 있다. 즉, 각 과목의 점수가 높으면 다른 과목의 점수도 높다는 것이다. 하지만 산점도 그래프의 분포 모양은 3 개의 그래프 모두 다른데, 가장 명확하게 직선 모양을 나타내는 것은 세 번째 그래프임을 알 수 있다. 따라서 4.2 에서 확인한 것과 같이 독해 점수와 작문 점수 간의 상관관계가 가장 높다는 것을 산점도 그래프를 통해서 확인할 수 있었다.

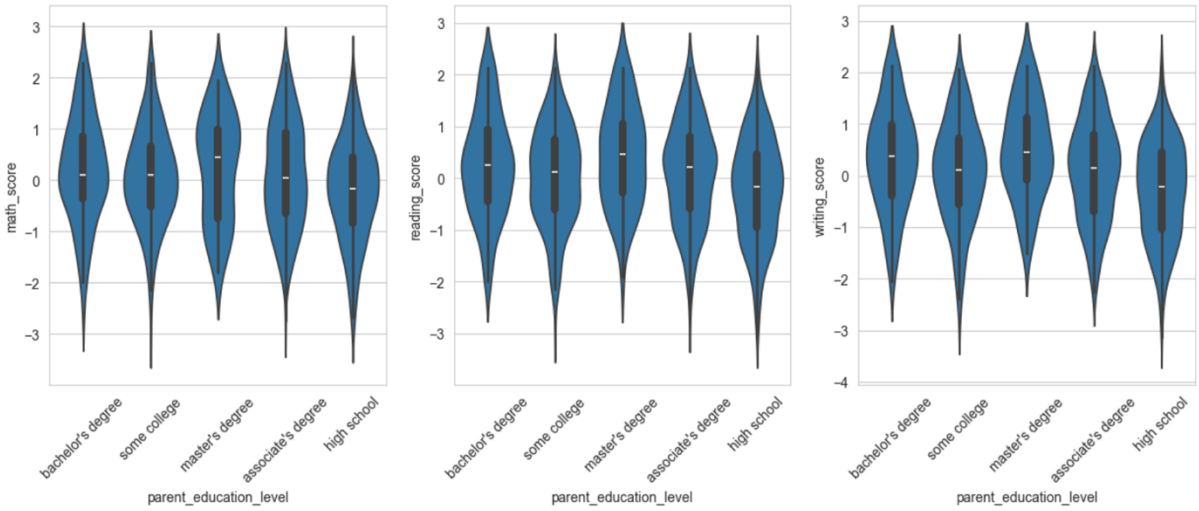
4.6. 부모의 학력 분포 확인



교육 사회학의 여러 연구에서는 부모의 사회 경제적 배경에 따라 자녀의 성적, 학벌이 다르다는 점을 밝혀냈다. 사회 경제적 배경이란 부모의 소득 뿐만 아니라 부모의 학력 등 다양한 요소가 포함된다. 본 데이터셋에는 부모의 학력이 포함되어 있었는데, 일반적으로 부모의 학력이 고학력일수록 자녀의 성적이 높으며, 명문대에 입학할 확률이 높다. 4.7, 4.8 에서 부모의 학력에 따른 성적 차이를 확인하기 전에 본 데이터셋의 부모의 학력 분포를 확인했다.

시각화 결과를 분석하면, 고등학교 졸업의 경우가 가장 높은 비율을 차지했고, 다음으로 대학 중퇴, 전문학사, 일반 학사, 석사 졸업 순으로 높은 비율을 차지했다.

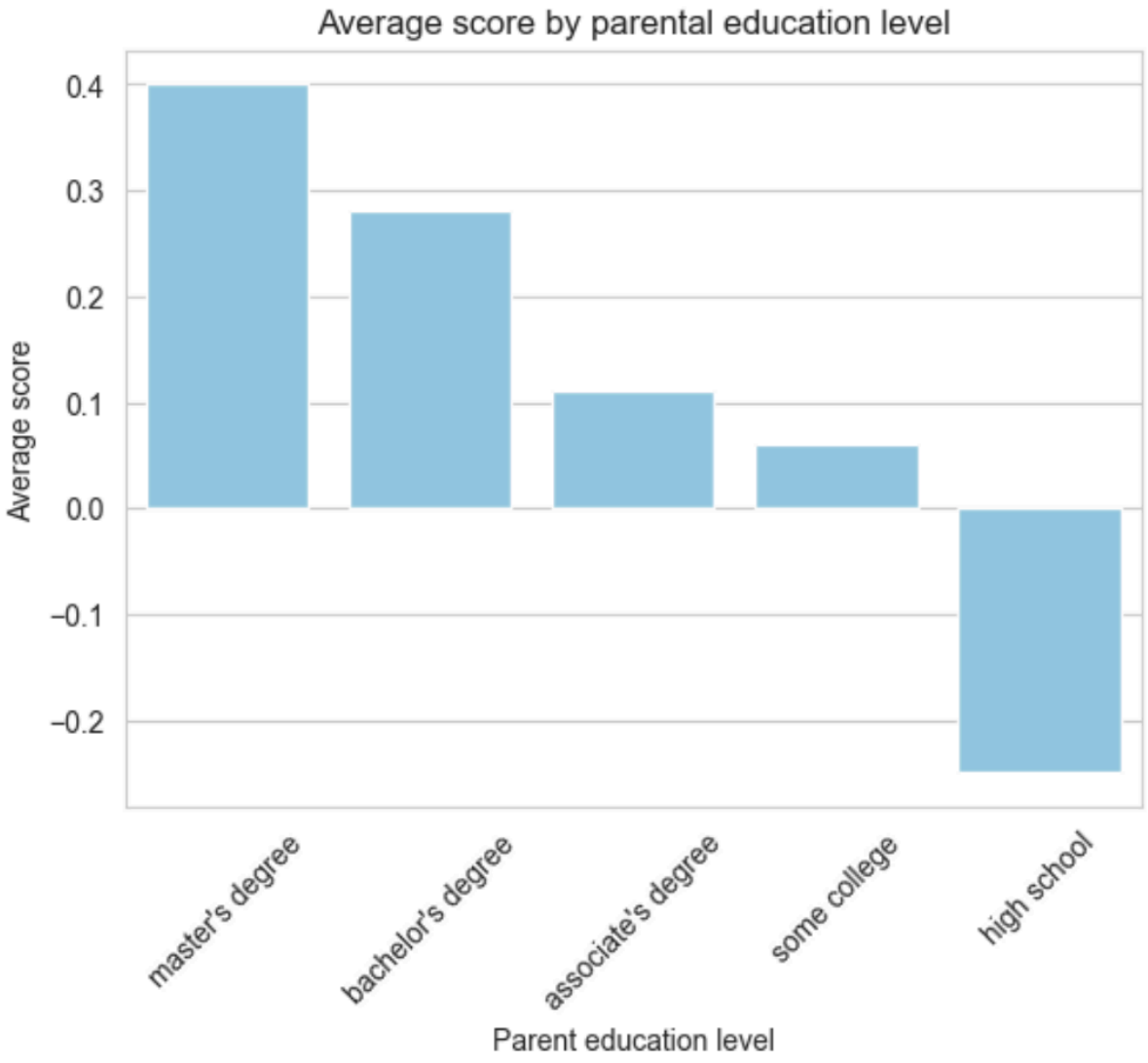
4.7. 부모의 학력에 기반한 각 과목별 시험 점수 바이올린 그래프



교육 사회학 연구들의 내용을 기반으로 부모의 학력이 낮을수록 자녀의 과목별 성적도 낮으며, 부모의 학력이 높을수록 자녀의 과목별 성적이 높을 것이라고 가설을 세우고 부모의 학력에 기반한 각 과목별 시험 점수를 바이올린 그래프로 시각화 하였다.

시각화 결과를 분석하면, 모든 과목에서 시험 점수의 최하위 분포를 보인 부모의 학력은 고등학교 졸업이었으며, 시험 점수가 가장 높은 분포가 발생하는 부모의 학력은 수학 과목을 제외한 독해 및 작문 과목에서는 모두 석사 졸업에서 발생했다. 또한 모든 과목에서 부모의 학력이 고졸인 경우, 다른 학력에 비해 전 과목에서 낮은 점수의 분포를 갖는다는 것을 확인할 수 있었고, 수학 과목에서는 석사 졸업보다 일반 학사 졸업 부모의 자녀가 더 높은 점수를 취득하는 경우도 있었다.

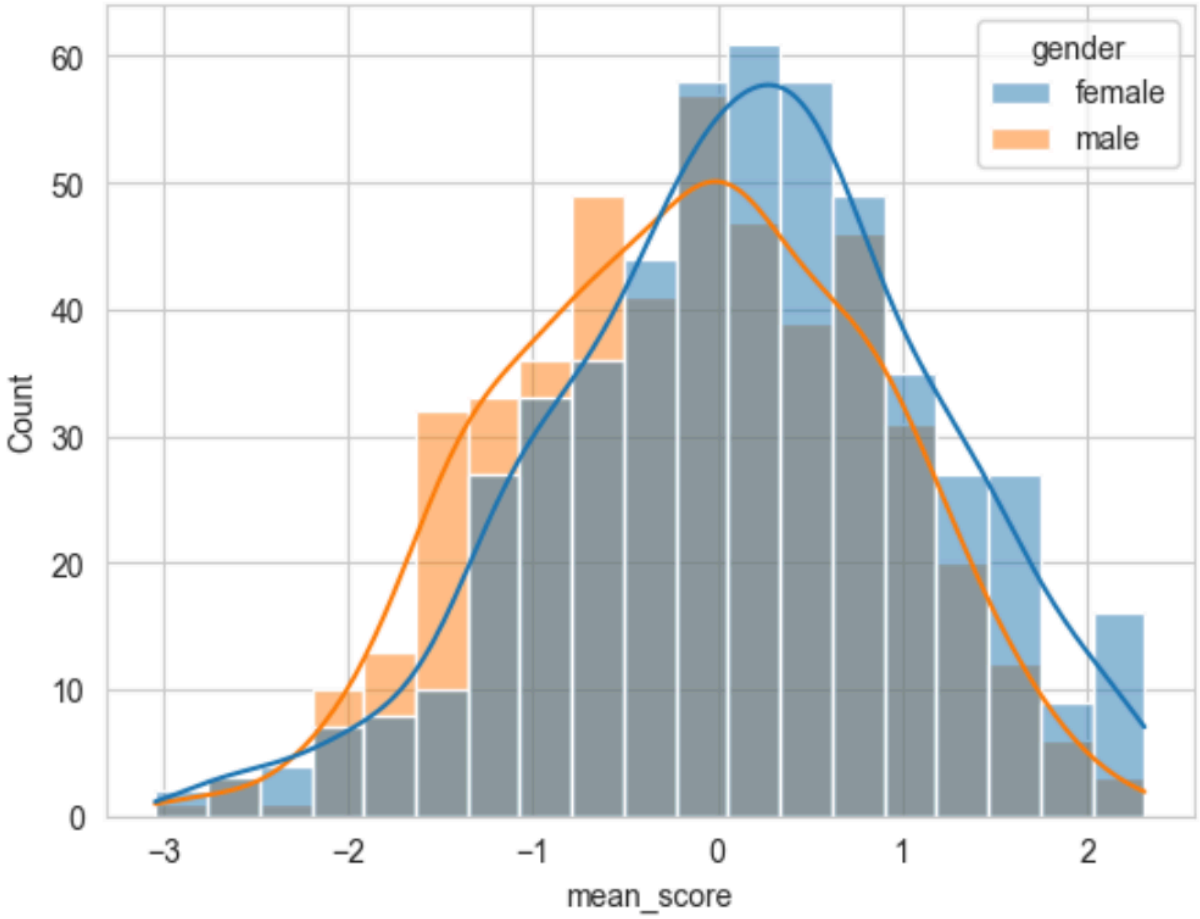
4.8. 부모의 학력별 평균 점수 분포



부모의 학력에 따른 과목별 시험 점수만으로는 확실한 분포 차이를 파악하기 어렵다. 따라서 데이터 전처리 과정에서 생성한 파생 변수인 학생의 평균 점수를 활용하여 부모의 학력별 평균 점수 분포를 시각화 하였다.

그 결과, 더욱 명확하게 부모의 학력에 따라 자녀의 시험 성적이 결정된다는 것을 확인할 수 있었다. y 축의 값은 데이터를 정규화해서 -1~1 사이의 값으로 변환되어 실제 시험 점수가 아닌 정규화된 값의 분포를 나타낸다. 먼저, 석사 졸업의 경우 자녀의 시험 평균 점수가 가장 높았다. 다음으로, 일반 학사, 전문학사, 대학 중퇴의 순서로 높았으며 타 집단과 다르게 고졸의 경우 훨씬 낮은 값을 확인할 수 있었다.

4.9. 성별에 따른 평균 점수 분포(커널 밀도 추정)



4.9 에서는 성별에 따른 학생의 시험 평균 점수를 히스토그램으로 나타냈으며 커널 밀도 추정을 통해 데이터 분포를 부드럽게 표현해서 시각화 하였다. 이때, 밀도는 데이터 개수 대비 비율이다.

시각화 결과를 분석하면, 여학생의 평균 점수 분포가 남학생보다 더 높은 평균 점수 쪽으로 치우쳐 있다는 것을 확인할 수 있었다. 즉, 여학생이 남학생보다 평균 점수가 더 높다는 것을 확인할 수 있었다. 또한 커널 밀도 추정 곡선을 통해 남학생보다 여학생의 평균 점수 분포가 더 넓고 평평한 것을 확인할 수 있었다.

5. References

[1] PSLeon, Student Study Performance, https://www.kaggle.com/datasets/bhavikjikadara/student-study-performance, 2024.04.13