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	: 2024-05-09
• 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. 평소 교육학 및 사회학에 관심이 있는 나로서는 과목 간의 상관관계, 성적과 부모의 학력 간의 연관성, 그리고 시험 직전의 식사가 시험 점수에 미치는 영향 등 다양한 변수 간의 관계를 탐색적 데이터 분석을 통해 밝혀내고, 향후 교육 관련 프로젝트에 반영하고자 해당 데이터 세트를 선정하게 되었다. 본 과제에서는 Seaborn 대신 인터랙티브한 시각화를 할 수 있는 Plotly 를 사용하여 데이터를 재분석하고자 한다.

III. 데이터 세트 출처

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

IV. 분석 도구

- Pandas, NumPy, Plotly, Matplotlib

동국대학교 dongguk university

2. 데이터 시각화

2.1. 해당 데이터셋의 남녀 비율 확인

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

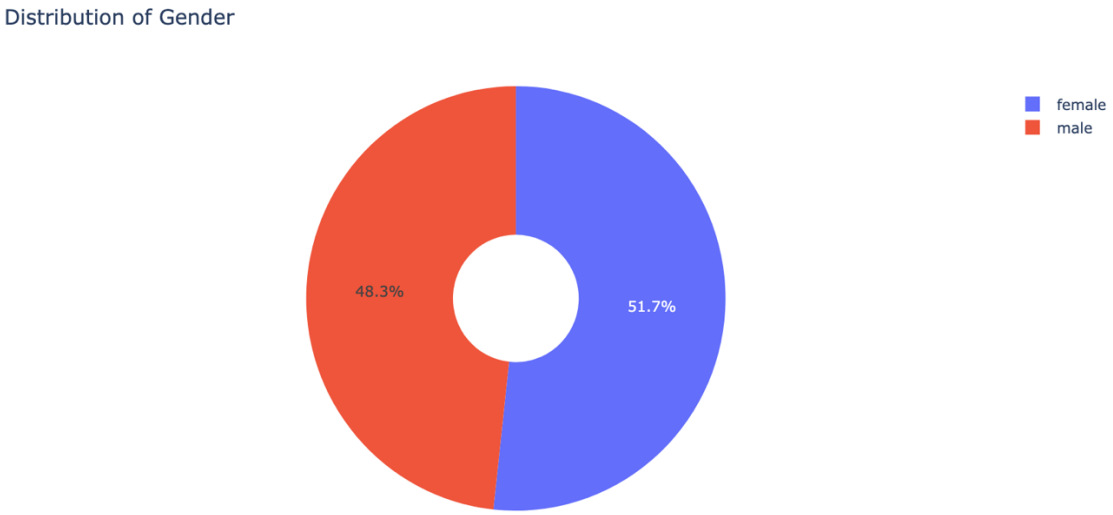
```
import plotly.graph_objects as go

# 각 column의 값에 대한 빈도수 계산
count_column = df['column_name'].value_counts()

# 파이 차트 생성
fig = go.Figure(data=[go.Pie(labels=count_column.index,
                             values=count_column,
                             textinfo='percent',
                             hoverinfo='label+percent',
                             hole=0.3)])

# 레이아웃 설정
fig.update_layout(title='Distribution of Column')

# 차트 표시
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 파이 차트를 데이터프레임의 특정 column 에 대해 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 column 을 지정하는 부분이 누락되어 있어서 필자가 원하는 것과 같이 gender 열에 대해서 파이차트로 시각화 하지 못한다는 이유로 코드를 그대로 사용할 수 없음. 따라서, 본 수업에서 배운 매개변수 활용 지식을 활용하여 labels 속성과 values 부분의 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
count_gender = df['gender'].value_counts()

fig = go.Figure(data=[go.Pie(labels=count_gender.index, values=count_gender, textinfo='percent',
                             hoverinfo='label+percent', hole=0.3)])

fig.update_layout(title='Distribution of Gender')

fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

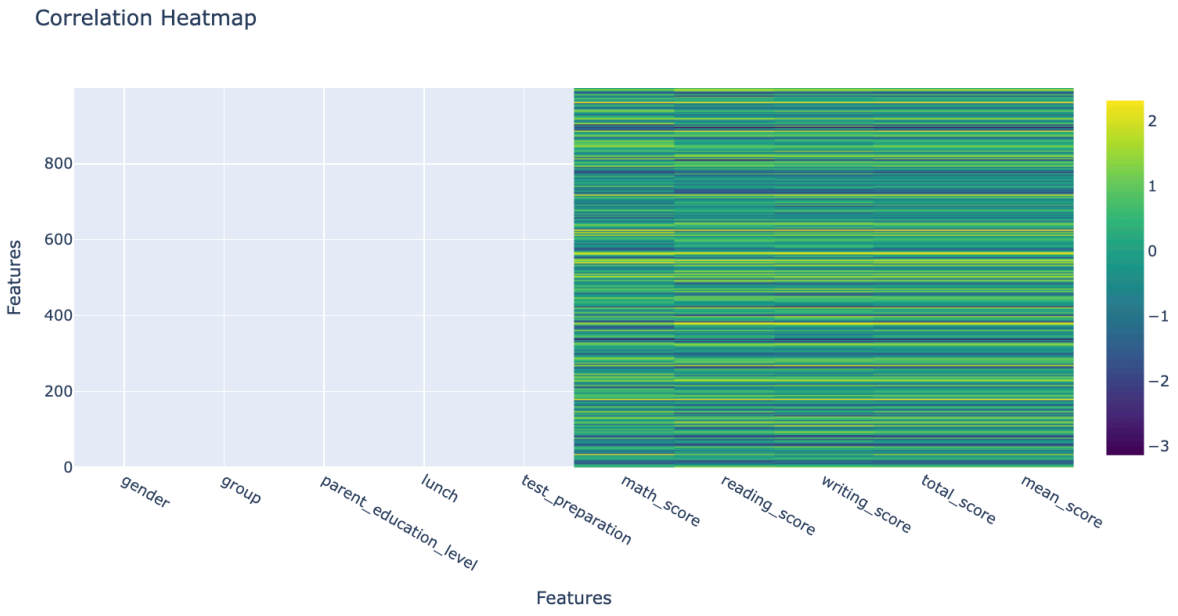
2.2. 수치형 데이터들의 상관관계 확인

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

```
import plotly.graph_objects as go

# 히트맵 그리기
fig = go.Figure(data=go.Heatmap(
                    z=df.values,
                    x=df.columns,
                    y=df.index,
                    colorscale='Viridis'))

fig.update_layout(title="Correlation Heatmap",
                  xaxis_title="Features",
                  yaxis_title="Features")

fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

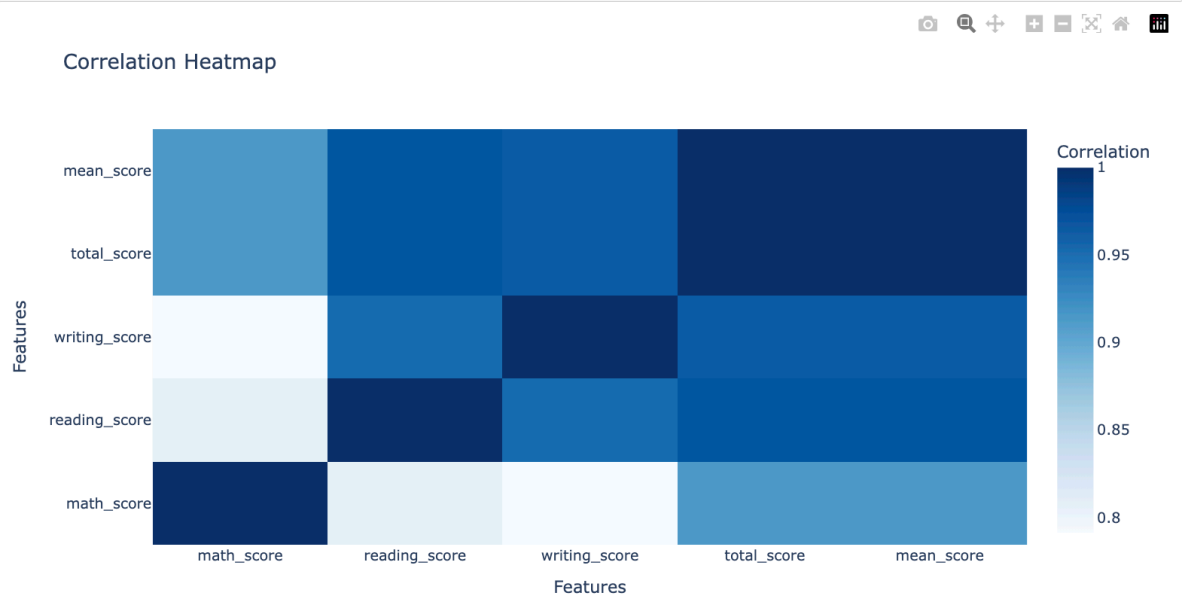
- 위 코드는 ChatGPT 가 제공한 상관관계를 히트맵으로 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드로 수행할 경우 데이터 프레임 전체 데이터에 대해 히트맵을 그리기 때문에 숫자형 데이터가 아닌 경우에도 아래와 같이 시각화 되므로 제대로 그려지지 않으며, 색상이 가독성이 떨어지기 때문에 코드를 그대로 사용할 수 없음.



따라서, 본 수업에서 배운 매개변수 활용 지식을 활용하여 numeric 데이터로만 이루어진 numeric_df 를 생성하고 이 데이터들로 상관행렬을 만들어 이에 대해 히트맵으로 시각화 하도록 코드를 수정함. 또한, 색상 스케일도 Blues 로 수정함.

4. 최종 코드 및 시각화 결과

```
numeric_df = df.select_dtypes(include=['number'])
correlation_matrix = numeric_df.corr()

fig = go.Figure(data=go.Heatmap(
                   z=correlation_matrix.values,
                   x=correlation_matrix.columns,
                   y=correlation_matrix.index,
                   colorscale='Blues',
                   colorbar=dict(title='Correlation')))

fig.update_layout(title='Correlation Heatmap',
                  xaxis=dict(title='Features'),
                  yaxis=dict(title='Features'))
fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

2.3. 성별에 따른 시험 준비 정도 확인

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

```
import plotly.express as px

# 카운트 플롯 그리기
fig = px.histogram(df, x='test_preparation', title='Count Plot of Test Preparation')

# 그래프 보여주기
fig.show()
```

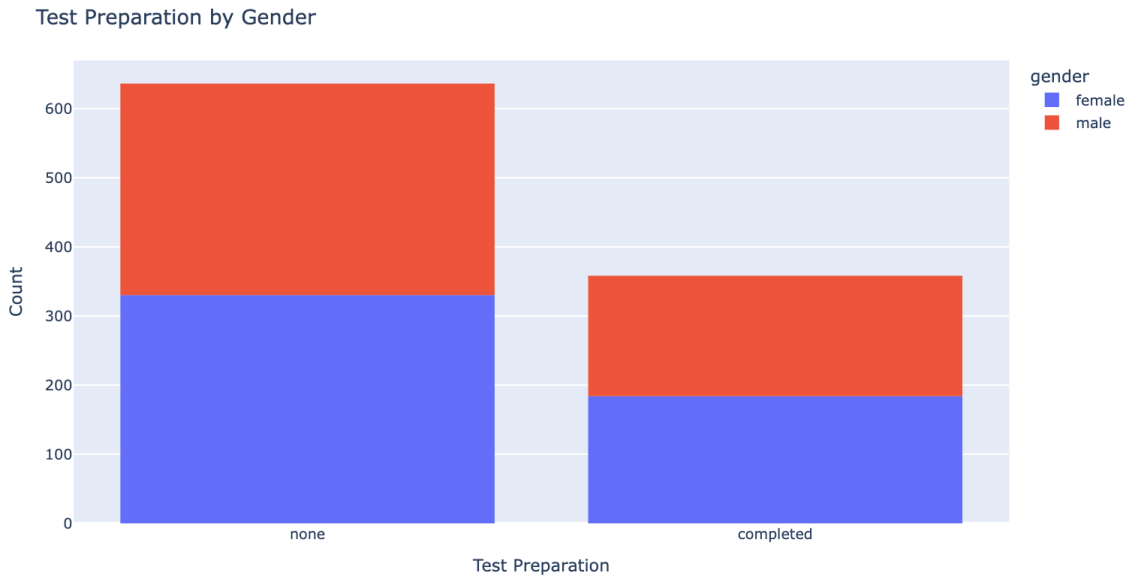
2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 성별에 따른 시험 준비 정도를 카운트 플롯으로 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 단순히 시험 준비 정도에 대해서만 카운트 플롯으로 나타내고 있어서 필자가 원하는 것과 같이 gender 열에 대해서 제대로 시각화 하지 못한다는 이유로 코드를 그대로 사용할 수 없음.

따라서, 본 수업에서 배운 매개변수 활용 지식을 활용하여 color 속성에 'gender'를 추가하는 방식으로 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
: fig = px.histogram(df, x='test_preparation', color='gender')
fig.update_layout(title='Test Preparation by Gender',
                  xaxis_title='Test Preparation',
                  yaxis_title='Count')
fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

4.4. 성별에 따른 시험 전 식사 여부 확인

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

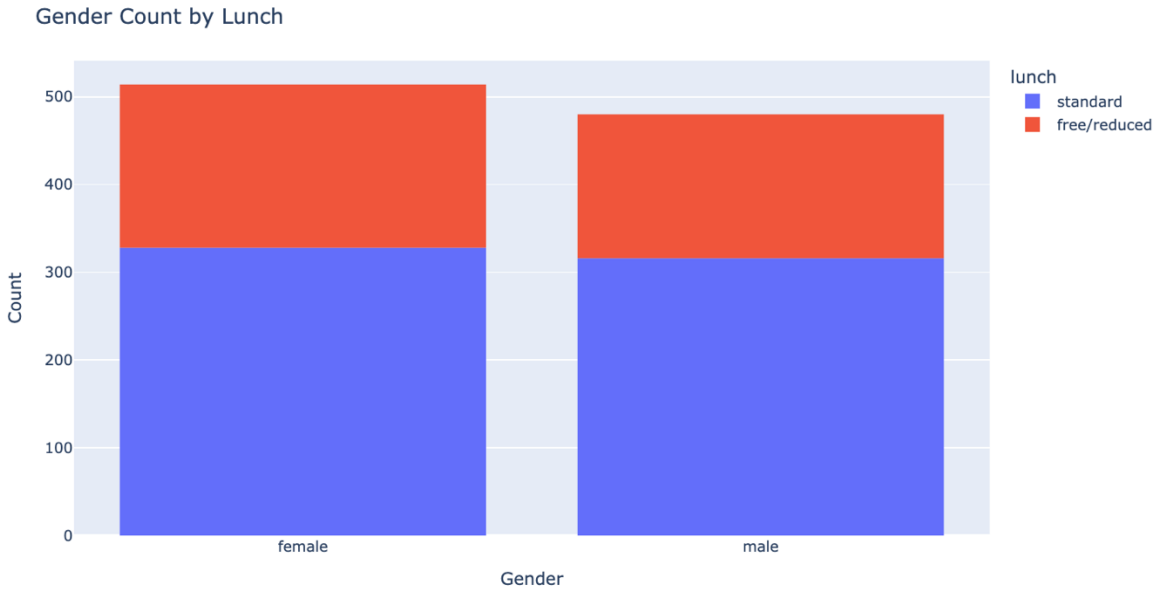
```
import plotly.express as px

# 카운트 플롯 그리기
fig = px.histogram(df, x='test_preparation', color='lunch', title='Count Plot of Test

# 그래프 보여주기
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 성별에 따른 시험 전 식사 여부를 카운트 플롯으로 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 이전의 질문과 연계되어 성별에 따른 시험 전 식사 여부가 아닌 식사 여부에 따른 시험 준비 정도에 대해 시각화를 하고 있다는 이유로 코드를 그대로 사용할 수 없음. 따라서, 본 수업에서 배운 매개변수 활용 지식을 활용하여 x 속성은 'gender'로, color 속성은 'lunch'로 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
fig = px.histogram(df, x='gender', color='lunch')
fig.update_layout(title='Gender Count by Lunch',
                  xaxis_title='Gender',
                  yaxis_title='Count')
fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

2.5. 각 과목별 산점도 그래프

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

```
import plotly.express as px

# 산점도 그래프 그리기
fig = px.scatter_3d(df, x='math_score', y='reading_score', z='writing_score', title=

# 그래프 보여주기
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 각 과목들의 점수에 대해 3d 산점도 그래프로 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 필자의 그래픽 환경에서는 3d 산점도가 나타나지 않았으며, 필자는 2d 산점도가 필요하다는 이유로 코드를 그대로 사용할 수 없음. 따라서, 본 수업에서 배운 plotly.graph_objects 의 add_trace() 함수와 go.Scatter()를 활용하여 각 산점도를 하나의 2d 그래프 상에 표현되도록 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
fig = go.Figure()

fig.add_trace(go.Scatter(x=df['reading_score'], y=df['math_score'], mode='markers', name='Reading vs Math'))
fig.add_trace(go.Scatter(x=df['writing_score'], y=df['math_score'], mode='markers', name='Writing vs Math'))
fig.add_trace(go.Scatter(x=df['reading_score'], y=df['writing_score'], mode='markers', name='Reading vs Writing'))

fig.update_layout(title='Score Distributions',
                  xaxis_title='Score',
                  yaxis_title='Score')

fig.show()
```

5. 시각화 결과 해석

- 첫 번째는 파란색 마커로, 독해 점수에 따른 수학 점수의 산점도 그래프이고, 두 번째는 빨간색 마커로, 작문 점수에 따른 수학 점수의 산점도 그래프이며, 마지막 그래프는 녹색 마커로, 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, 2.2 에서 확인한 것과 같이 독해 점수와 작문 점수 간의 상관관계가 가장 높다는 것을 산점도 그래프를 통해서 확인할 수 있었다.

동국대학교 dongguk university

2.6. 부모의 학력 분포 확인

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

python Copy code

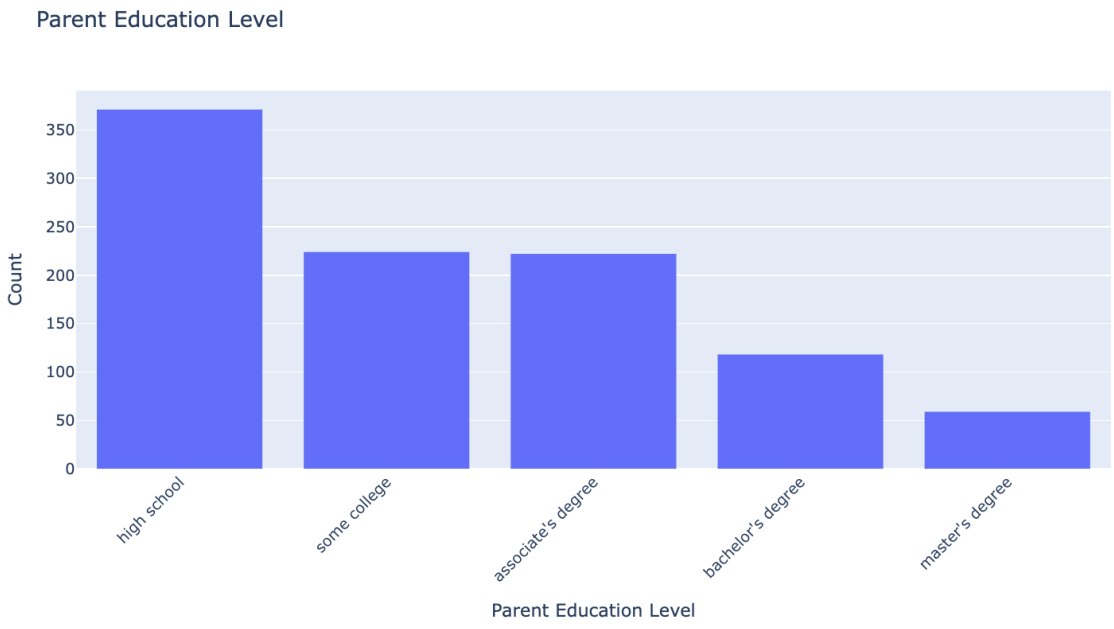
```
import plotly.express as px

fig = px.box(df, x='parent_education_level', title='Distribution of Parent Education
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 부모의 학력 분포를 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 필자가 기존에 구현한 것과 같이 카운트 플롯 형태가 아닌 상자 그림 형태로 제공하였기 때문에 코드를 그대로 사용할 수 없음. 따라서, 본 수업에서 배운 plotly.express 의 histogram() 함수를 활용하도록 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
fig = px.histogram(df, x='parent_education_level', title='Parent Education Level')
fig.update_layout(xaxis_title='Parent Education Level',
                  yaxis_title='Count',
                  xaxis={'categoryorder':'total descending'},
                  xaxis_tickangle=-45)
fig.show()
```



5. 시각화 결과 해석

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

동국대학교 dongguk university

2.7. 부모의 학력에 기반한 각 과목별 시험 점수 바이올린 플롯

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

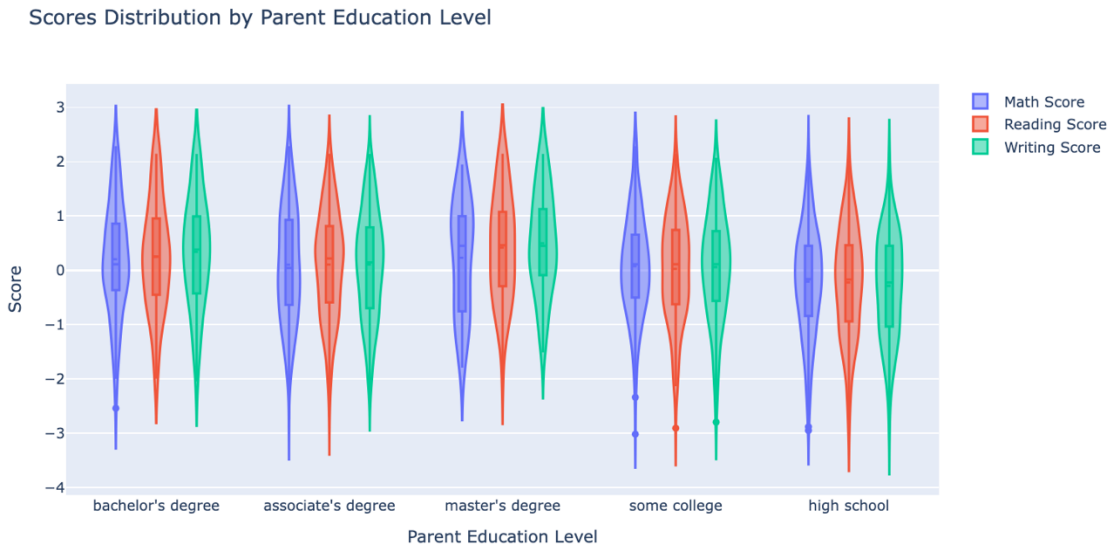
```
import plotly.express as px

fig = px.violin(df, x='parent_education_level', y='math_score', title='Math Score by
fig.show()

fig = px.violin(df, x='parent_education_level', y='reading_score', title='Reading Sco
fig.show()

fig = px.violin(df, x='parent_education_level', y='writing_score', title='Writing Sco
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 부모의 학력에 기반한 각 과목별 시험 점수를 바이올린 플롯으로 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 각 과목별 바이올린 플롯을 개별적으로 시각화 하기 때문에 인터랙티브한 시각화가 가능한 plotly 의 장점을 제대로 살릴 수 없기 때문에 코드를 그대로 사용할 수 없음. 따라서, 본 수업에서 배운 add_trace()를 사용하여 하나의 그래프에 중첩하여 나타내도록 코드를 수정함. 이렇게 함으로써 하나의 그래프에 여러 과목의 점수 분포를 비교할 수 있음.

4. 최종 코드 및 시각화 결과

```
fig = go.Figure()

fig.add_trace(go.Violin(x=df['parent_education_level'], y=df['math_score'], name='Math Score', box_visible=True, mea
fig.add_trace(go.Violin(x=df['parent_education_level'], y=df['reading_score'], name='Reading Score', box_visible=Tru
fig.add_trace(go.Violin(x=df['parent_education_level'], y=df['writing_score'], name='Writing Score', box_visible=Tru

fig.update_layout(title='Scores Distribution by Parent Education Level',
                  xaxis_title='Parent Education Level',
                  yaxis_title='Score',
                  xaxis={'categoryorder':'total descending'},
                  violinmode='group')

fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

2.8. 부모의 학력별 평균 점수 분포

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

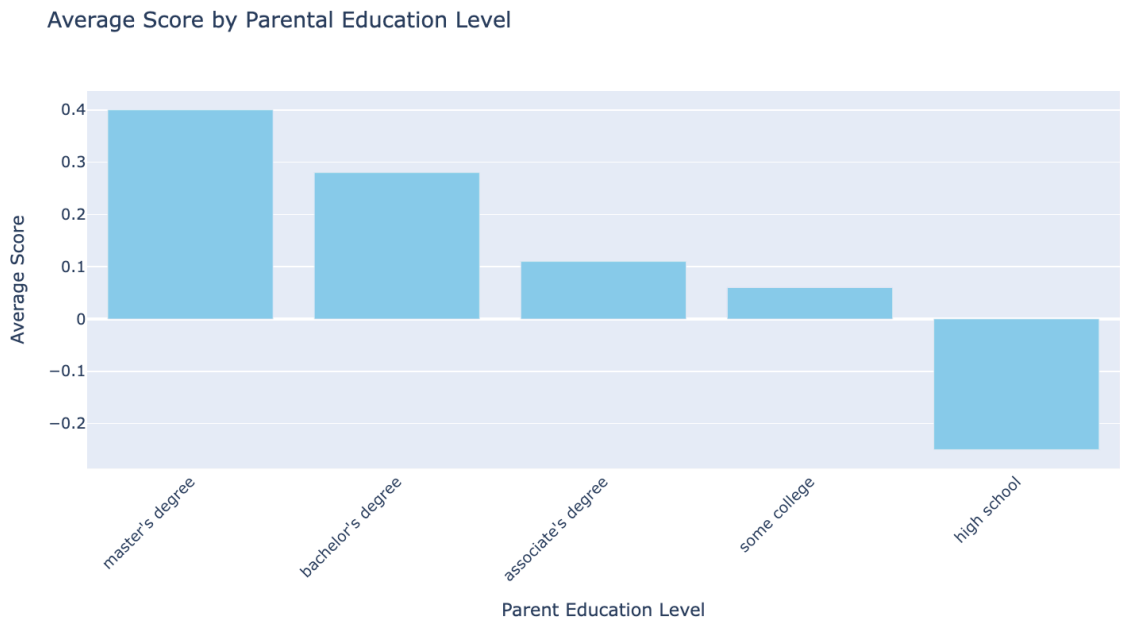
```
import plotly.graph_objects as go

# 부모의 학력별 각 과목의 평균 점수 계산
mean_scores = df.groupby('parent_education_level')[['math_score', 'reading_score', 'w

# 바이올린 그래프 생성
fig = go.Figure()

for col in mean_scores.columns:
    fig.add_trace(go.Violin(x=df['parent_education_level'], y=mean_scores[col], name=

# 그래프 레이아웃 설정
fig.update_layout(title='Average Scores Distribution by Parent Education Level',
                  xaxis_title='Parent Education Level',
                  yaxis_title='Average Score',
                  xaxis={'categoryorder':'total descending'},
                  violinmode='group')

# 그래프 출력
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 부모의 학력별 전체 과목의 시험 점수를 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 필자가 기존에 구현한 바 플롯이 아닌 바이올린 플롯으로 시각화 하기 때문에 코드를 그대로 사용할 수 없음. 따라서, go.Violin() 대신 본 수업에서 배운 go.Bar()로 코드를 수정하였으며 violinmode 인자값도 삭제함.

4. 최종 코드 및 시각화 결과

```
mean_scores = df.groupby('parent_education_level')['mean_score'].mean().round(2).sort_values(ascending=False)

fig = go.Figure(data=go.Bar(x=mean_scores.index, y=mean_scores.values, marker_color='skyblue'))

fig.update_layout(xaxis=dict(title='Parent Education Level'),
                  yaxis=dict(title='Average Score'),
                  xaxis_tickangle=-45,
                  title='Average Score by Parental Education Level')

fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

2.9. 성별에 따른 평균 점수 분포

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

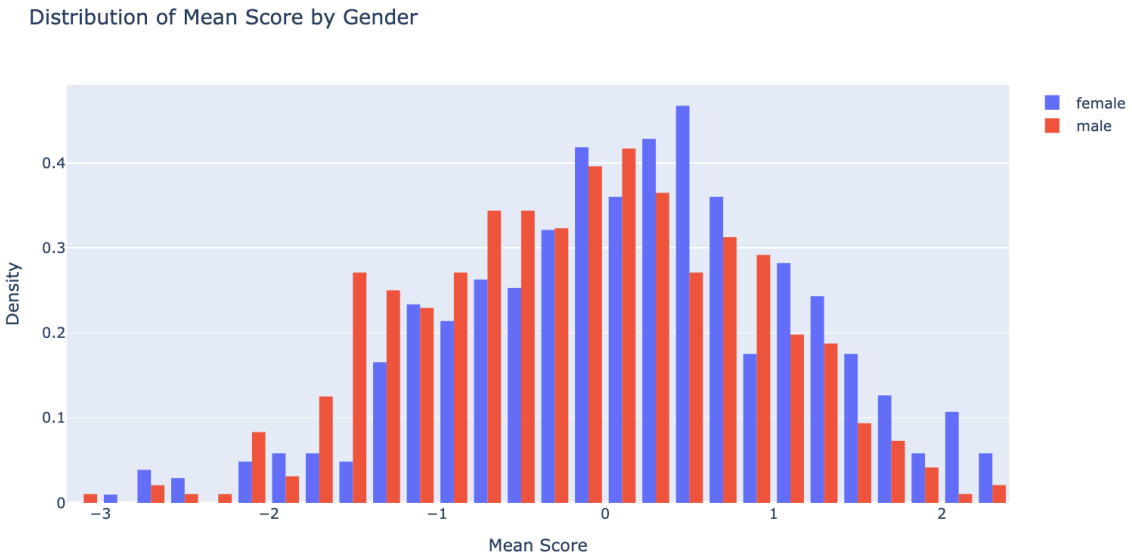
```
import plotly.graph_objects as go

# 성별별 각 과목의 평균 점수 계산
mean_scores_gender = df.groupby('gender')[['math_score', 'reading_score', 'writing_sc

# 바이올린 그래프 생성
fig = go.Figure()

for col in mean_scores_gender.columns:
    fig.add_trace(go.Violin(x=df['gender'], y=mean_scores_gender[col], name=col, box_

# 그래프 레이아웃 설정
fig.update_layout(title='Average Scores Distribution by Gender',
                  xaxis_title='Gender',
                  yaxis_title='Average Score',
                  violinmode='group')

# 그래프 출력
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 성별에 따른 평균 점수 분포를 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 필자가 기존에 구현한 히스토그램이 아닌 바이올린 플롯으로 시각화 하기 때문에 코드를 그대로 사용할 수 없음. 따라서, go.Violin() 대신 본 수업에서 배운 go.Histogram()으로 코드를 수정하였으며 violinmode 인자값도 삭제함.

4. 최종 코드 및 시각화 결과

```
fig = go.Figure()

for gender_value in df['gender'].unique():
    gender_data = df[df['gender'] == gender_value]['mean_score']
    fig.add_trace(go.Histogram(x=gender_data, name=gender_value, histnorm='probability density', nbinsx=30))

fig.update_layout(title='Distribution of Mean Score by Gender',
                  xaxis_title='Mean Score',
                  yaxis_title='Density')

fig.show()
```



5. 시각화 결과 해석

- 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.

동국대학교 dongguk university

2.10. 성별에 따른 수학 점수 분포(상자그림)

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

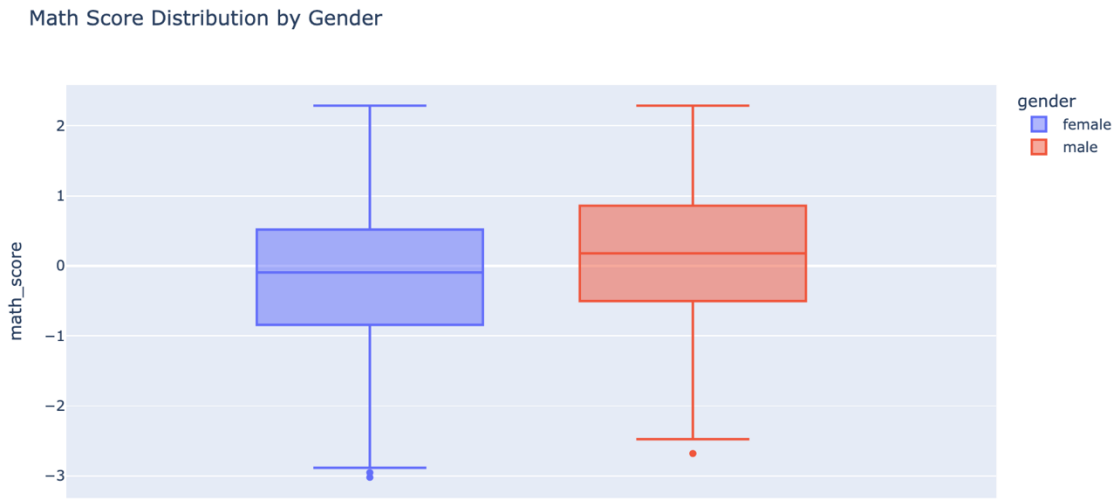
```
import plotly.graph_objects as go

# 성별에 따른 수학 점수 데이터
math_scores_male = male['math_score']
math_scores_female = female['math_score']

# 상자 그림 생성
fig = go.Figure()

fig.add_trace(go.Box(y=math_scores_male, name='Male', boxmean=True))
fig.add_trace(go.Box(y=math_scores_female, name='Female', boxmean=True))

# 그래프 레이아웃 설정
fig.update_layout(title='Math Score Distribution by Gender',
                  yaxis_title='Math Score',
                  boxmode='group')

# 그래프 출력
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 성별에 따른 수학 점수 분포를 상자그림으로 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 성별에 따른 수학 점수 데이터 지정이 잘못되어 데이터를 제대로 불러올 수 없기 때문에 코드를 그대로 사용할 수 없음. 따라서, 본 수업에서 배운 Data Manipulation 파트의 코드를 참조하여 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
math_scores_male = df[df['gender'] == 'male']['math_score']
math_scores_female = df[df['gender'] == 'female']['math_score']

fig = px.box(df, y='math_score', color='gender', title='Math Score Distribution by Gender')

fig.show()
```



5. 시각화 결과 해석

- 수학 점수의 경우, 남성이 여성보다 더 높은 점수대에 분포한다는 것을 확인할 수 있으며, 점수가 최하위에 분포한 학생도 남성보다 여성이 다수라는 것을 확인할 수 있었다.

동국대학교 dongguk university

4.11. 성별에 따른 독해 점수 분포(바이올린 플롯)

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

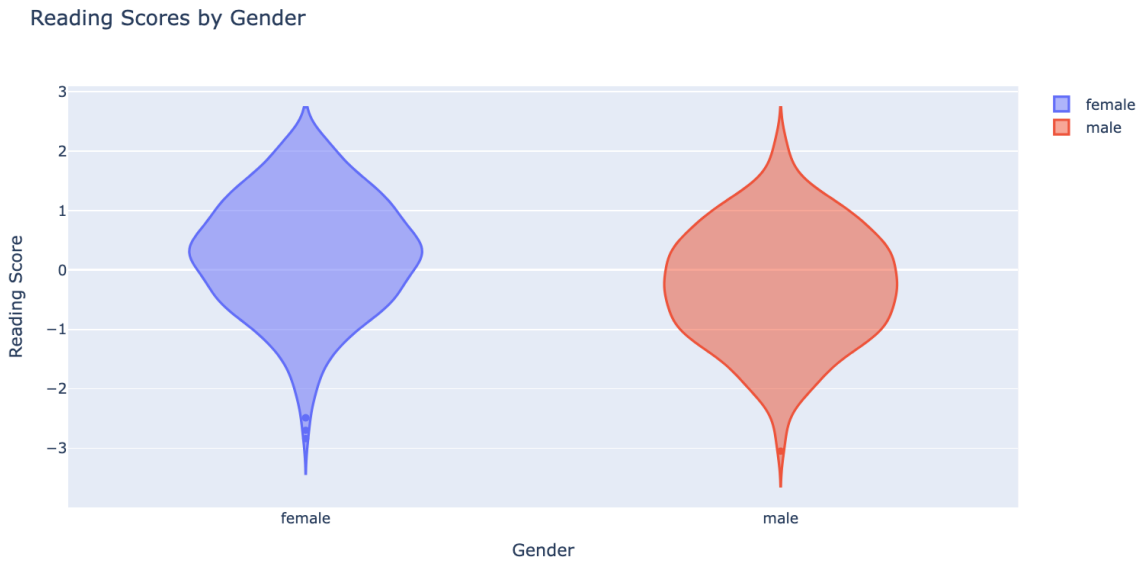
```
import plotly.graph_objects as go

# 성별에 따른 독해 점수 데이터
reading_scores_male = df[df['gender'] == 'male']['reading_score']
reading_scores_female = df[df['gender'] == 'female']['reading_score']

# 상자 그림 생성
fig = go.Figure()

fig.add_trace(go.Box(y=reading_scores_male, name='Male', boxmean=True))
fig.add_trace(go.Box(y=reading_scores_female, name='Female', boxmean=True))

# 그래프 레이아웃 설정
fig.update_layout(title='Reading Score Distribution by Gender',
                  yaxis_title='Reading Score',
                  boxmode='group')

# 그래프 출력
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 부모의 학력별 전체 과목의 시험 점수를 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 바이올린 플롯이 아닌 상자그림으로 시각화 하기 때문에 코드를 그대로 사용할 수 없음. 따라서, go.Box() 대신 본 수업에서 배운 go.Violin()으로 코드를 수정함.

4. 최종 코드 및 시각화 결과

```python
# Violin plot for reading scores by gender
violin_plot_reading = go.Figure()
for gender_value in df['gender'].unique():
    gender_data = df[df['gender'] == gender_value]['reading_score']
    violin_plot_reading.add_trace(go.Violin(y=gender_data, name=gender_value))
violin_plot_reading.update_layout(title='Reading Scores by Gender',
                                  xaxis_title='Gender',
                                  yaxis_title='Reading Score')
```



5. 시각화 결과 해석

- 독해 점수의 경우, 남녀 모두 비슷한 분포를 보였지만, 남성의 경우 더 고득점 비율이 높았다. 또한, 여성의 경우 대다수가 평균 점수대에 분포되어 있음을 확인할 수 있었다.

동국대학교 dongguk university

4.12. 부모의 학력에 따른 각 과목별 평균 점수 분포

1. 중간 발표 때 구현한 기능을 ChatGPT 로 구현한 코드

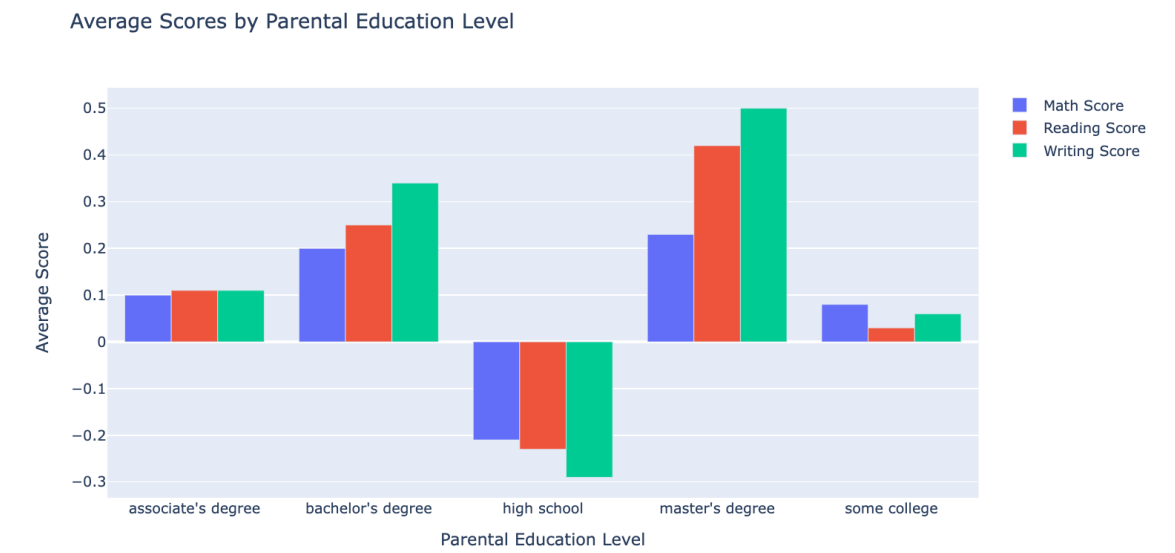
```
import plotly.graph_objects as go

# 부모의 학력에 따른 각 과목별 평균 점수 데이터
mean_math_scores = df.groupby('parent_education_level')['math_score'].mean()
mean_reading_scores = df.groupby('parent_education_level')['reading_score'].mean()
mean_writing_scores = df.groupby('parent_education_level')['writing_score'].mean()

# 상자 그림 생성
fig = go.Figure()

fig.add_trace(go.Box(y=df['math_score'], x=df['parent_education_level'], name='Math
fig.add_trace(go.Box(y=df['reading_score'], x=df['parent_education_level'], name='Re
fig.add_trace(go.Box(y=df['writing_score'], x=df['parent_education_level'], name='Wr

# 그래프 레이아웃 설정
fig.update_layout(title='Score Distributions by Parent Education Level',
                  xaxis_title='Parent Education Level',
                  yaxis_title='Score',
                  boxmode='group')

# 그래프 출력
fig.show()
```

2 & 3. 위 코드를 그대로 사용할 수 없는 이유와 이를 해결하기 위해 본 수업에서 배운 지식을 활용한 부분

- 위 코드는 ChatGPT 가 제공한 부모의 학력에 따른 각 과목별 평균 점수 분포를 시각화 하는 코드임. 그러나 ChatGPT 로 구현한 코드는 바 플롯이 아닌 상자그림으로 시각화 하기 때문에 명확한 분포 비교가 어렵다는 점에서 코드를 그대로 사용할 수 없음. 따라서, go.Box() 대신 본 수업에서 배운 go.Bar()로 코드를 수정함.

4. 최종 코드 및 시각화 결과

```
# Bar plot for average scores by parental education level
mean_scores_parent_education = df.groupby('parent_education_level')[['math_score', 'reading_score', 'writing_score']
bar_plot_parent_education = go.Figure(data=[
    go.Bar(name='Math Score', x=mean_scores_parent_education.index, y=mean_scores_parent_education['math_score']),
    go.Bar(name='Reading Score', x=mean_scores_parent_education.index, y=mean_scores_parent_education['reading_score
    go.Bar(name='Writing Score', x=mean_scores_parent_education.index, y=mean_scores_parent_education['writing_score
])
bar_plot_parent_education.update_layout(barmode='group',
                                        title='Average Scores by Parental Education Level',
                                        xaxis_title='Parental Education Level',
                                        yaxis_title='Average Score')
```



5. 시각화 결과 해석

- 분포를 확인한 결과, 고졸 부모의 경우 모든 과목에서 자녀의 성적이 낮았으며, 반면에 학사나 석사 졸업의 경우 자녀의 성적이 모든 과목에서 타 학력보다 높았음을 시각화를 통해 확인할 수 있었다.

3. 느낀점

중간 프로젝트와 이번 과제를 통해 다양한 데이터 분석 기법과 Seaborn 및 Plotly 와 같은 시각화 도구를 활용하여 실제 교육 관련 데이터를 깊이 있게 분석해볼 수 있었다. 중간 프로젝트 때에는 평소에 자주 사용하던 Seaborn 을 사용했지만, 이번 과제에서는 Seborn 대신 Plotly 를 사용하여 인터랙티브한 시각화를 할 수 있게 되어 데이터를 더 다양한 각도에서 살펴볼 수 있었다. 특히, 각 변수 간의 상관관계와 패턴을 시각적으로 탐색하고 이를 통해 교육 관련 요인들이 학생들의 성적에 미치는 영향을 조사하는 것은 매우 흥미로웠다. 또한, 부모의 학력과 자녀의 성적 간의 연관성을 발견했을 때와 같이 탐색적 데이터 분석을 통해 타 연구에서 밝혀진 내용이 맞다는 것을 확인했을 때 가장 큰 성취감을 느꼈다.

그리고 대학교에서 처음으로 ChatGPT 를 활용하여 코드를 작성하고 수정하는 과제를 진행했는데, 이 과정 동안 수행한 내용이 AI 를 활용한 학습에 매우 유용했다. 특히, 데이터 시각화를 위한 Plotly 와 관련된 지식을 습득하는 과정에서 ChatGPT 가 제공한 코드를 토대로 새로운 시각화 기법을 적용하고 수정하는 연습을 할 수 있었으며 ChatGPT 를 활용하여 코드를 작성하는 과정에서 Plotly 를 활용하기 위해 필요한 문법과 함수에 대한 이해도를 높일 수 있었고, AI 를 활용한 데이터 분석 및 시각화에 대해 새로운 경험을 쌓을 수 있었다.

마지막으로, 이번 프로젝트에서 교육 데이터를 분석하여 캐글에 공유하여 15 개의 upvote 를 받았다. 다음에도 실제 나의 데이터 분석 코드를 캐글에 공유하여 더 높은 메달을 획득하고 싶으며, 이번 경험을 토대로 향후 데이터 분석 프로젝트에 보다 자신 있게 참여하고 싶다.

4. References

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