In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

- Importando e entendendo base de dados

In [2]:
```
#Ler aquivo CSV
df = pd.read_csv(r"data\student_habits_performance.csv")
```

In [3]:
```
#Visualizar dados
df
```

Out[3]:

	student_id	age	gender	study_hours_per_day	social_media_hours	netflix_hours	p
0	S1000	23	Female	0.0	1.2	1.1	
1	S1001	20	Female	6.9	2.8	2.3	
2	S1002	21	Male	1.4	3.1	1.3	
3	S1003	23	Female	1.0	3.9	1.0	
4	S1004	19	Female	5.0	4.4	0.5	
...	...	...	...	...	...	...	
995	S1995	21	Female	2.6	0.5	1.6	
996	S1996	17	Female	2.9	1.0	2.4	
997	S1997	20	Male	3.0	2.6	1.3	
998	S1998	24	Male	5.4	4.1	1.1	
999	S1999	19	Female	4.3	2.9	1.9	

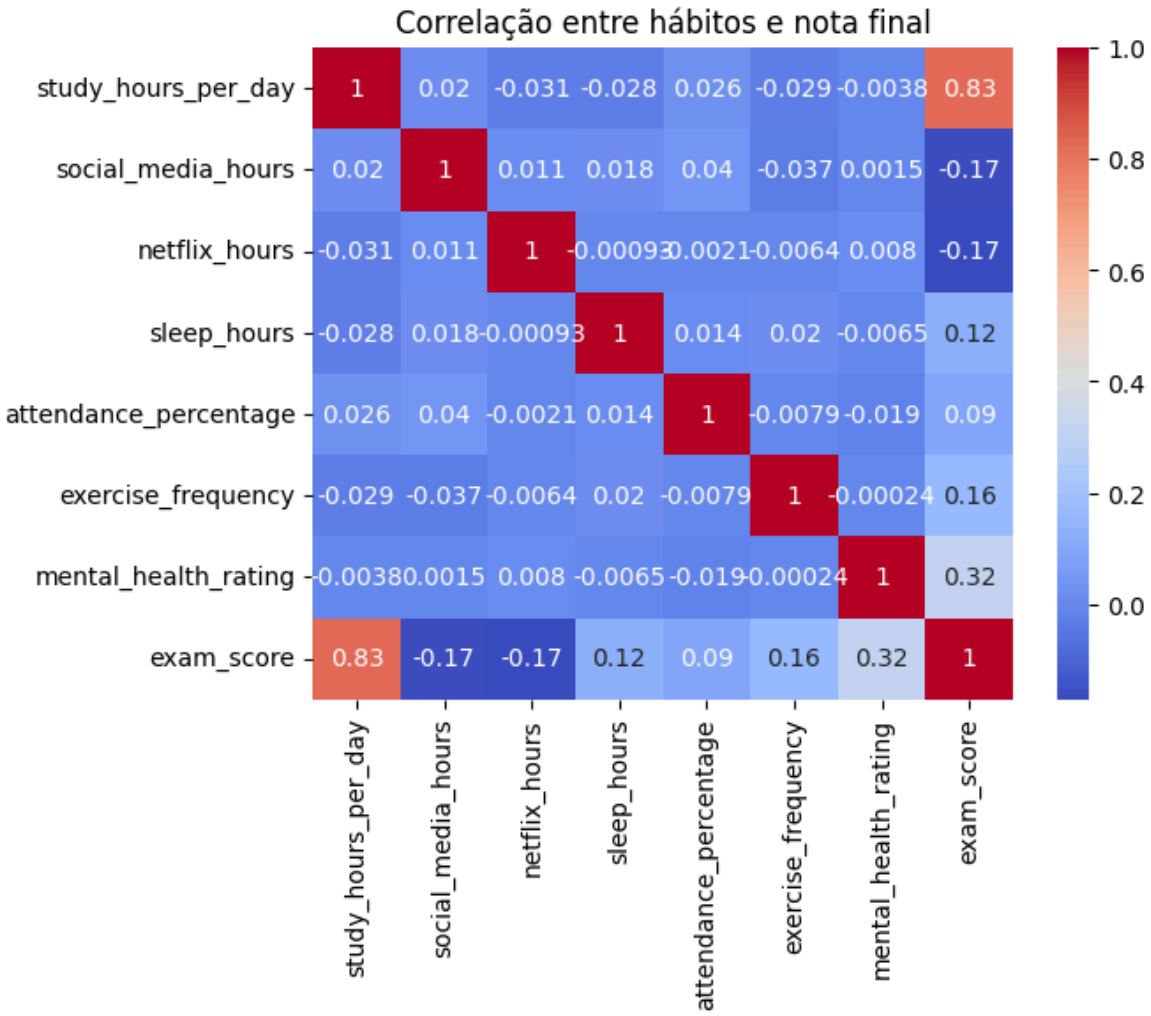
1000 rows × 16 columns

- Quais hábitos impactam mais o desempenho dos alunos?

In [4]:
```
#Tipos de dados

df.info()
```

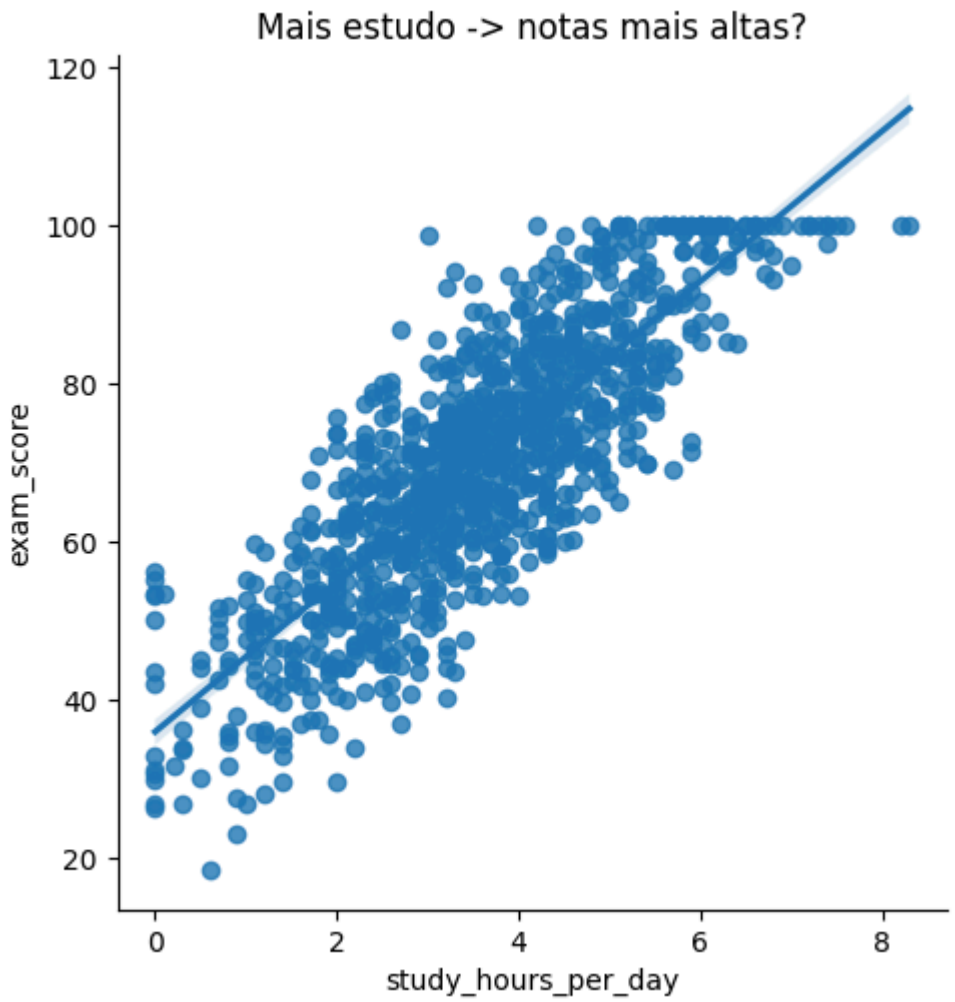
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   student_id                     1000 non-null   object 
 1   age                            1000 non-null   int64  
 2   gender                         1000 non-null   object 
 3   study_hours_per_day            1000 non-null   float64
 4   social_media_hours             1000 non-null   float64
 5   netflix_hours                  1000 non-null   float64
 6   part_time_job                  1000 non-null   object 
 7   attendance_percentage          1000 non-null   float64
 8   sleep_hours                    1000 non-null   float64
 9   diet_quality                   1000 non-null   object 
 10  exercise_frequency             1000 non-null   int64  
 11  parental_education_level       909 non-null    object 
 12  internet_quality               1000 non-null   object 
 13  mental_health_rating           1000 non-null   int64  
 14  extracurricular_participation  1000 non-null   object 
 15  exam_score                     1000 non-null   float64
dtypes: float64(6), int64(3), object(7)
memory usage: 125.1+ KB
```

In [5]:
```python
#Colunas numéricas
cols = [
    "study_hours_per_day",
    "social_media_hours",
    "netflix_hours",
    "sleep_hours",
    "attendance_percentage",
    "exercise_frequency",
    "mental_health_rating",
    "exam_score",
]

# Plotar mapa de calor (heatmap)
sns.heatmap(df[cols].corr(), annot=True, cmap="coolwarm")
plt.title("Correlação entre hábitos e nota final")
plt.show()
```



- Alunos que estudam mais tem melhor desempenho?

In [6]:
```
# Gráfico de dispersao com linha de regressao
# x="study_hours_per_day" / y="exam_score"
sns.lmplot(data=df, x="study_hours_per_day", y="exam_score")
plt.title("Mais estudo -> notas mais altas?")
plt.show()
```

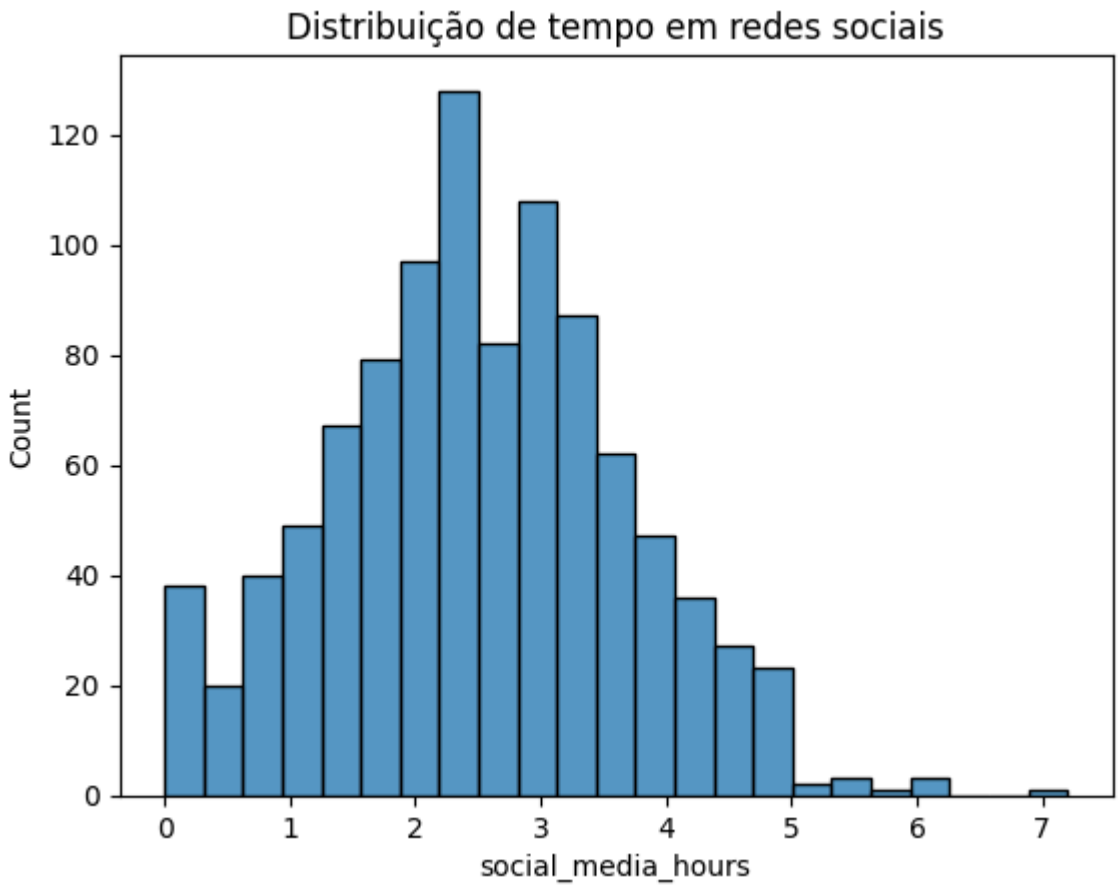


In [7]:
```python
#Comparado médias: quem estuda >5h x <2h
filtro_Estudo_alto = df["study_hours_per_day"] > 5
filtro_estudo_baixo = df["study_hours_per_day"] < 2

grupo_estudo_alto = df[filtro_Estudo_alto]["exam_score"]
grupo_estudo_baixo = df[filtro_estudo_baixo]["exam_score"]

print("Média notas (estuda > 5h):", grupo_estudo_alto.mean())
print("Média notas (estuda < 2h):", grupo_estudo_baixo.mean())
```

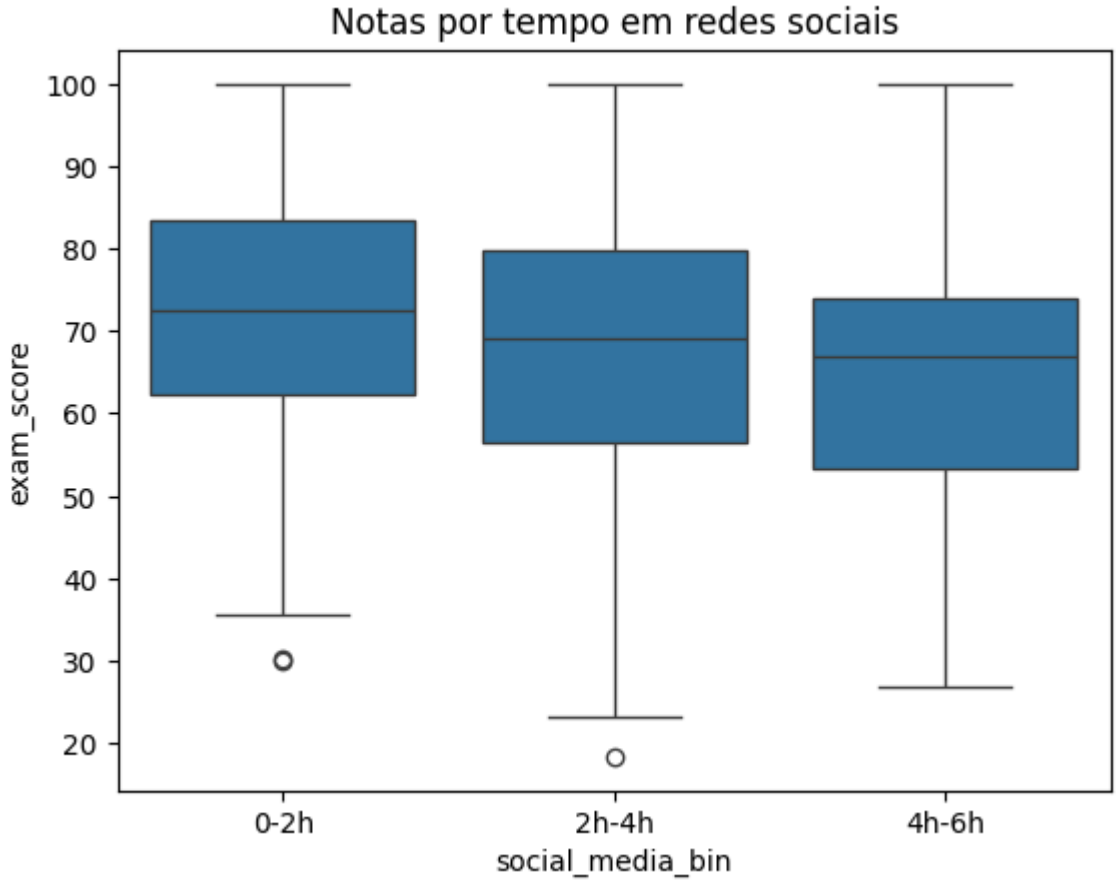
```
Média notas (estuda > 5h): 90.79419354838709
Média notas (estuda < 2h): 45.56390977443609
```

- O tempo gasto em redes sociais afeta o desempenho dos alunos?

In [8]:
```python
# Redes sociais: distribuição geral (Histograma)
# x="social_media_hours"

sns.histplot(data=df, x="social_media_hours")
plt.title("Distribuição de tempo em redes sociais")
plt.show()
```

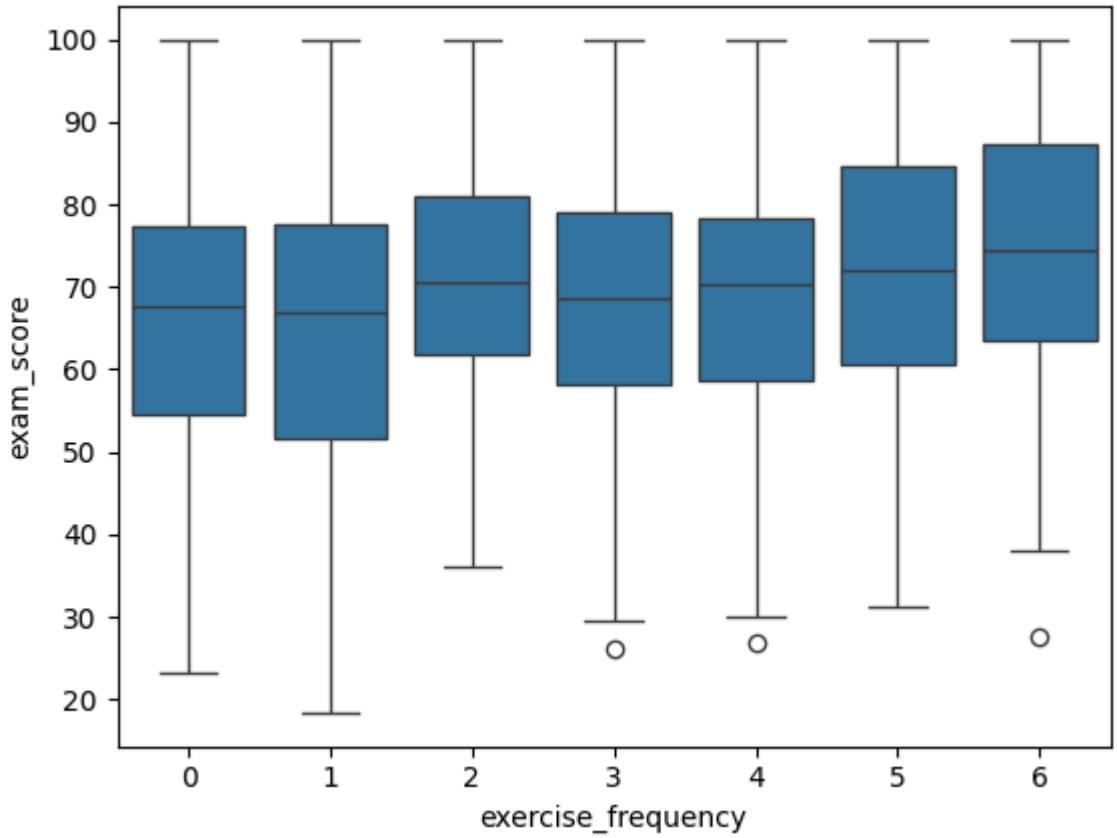


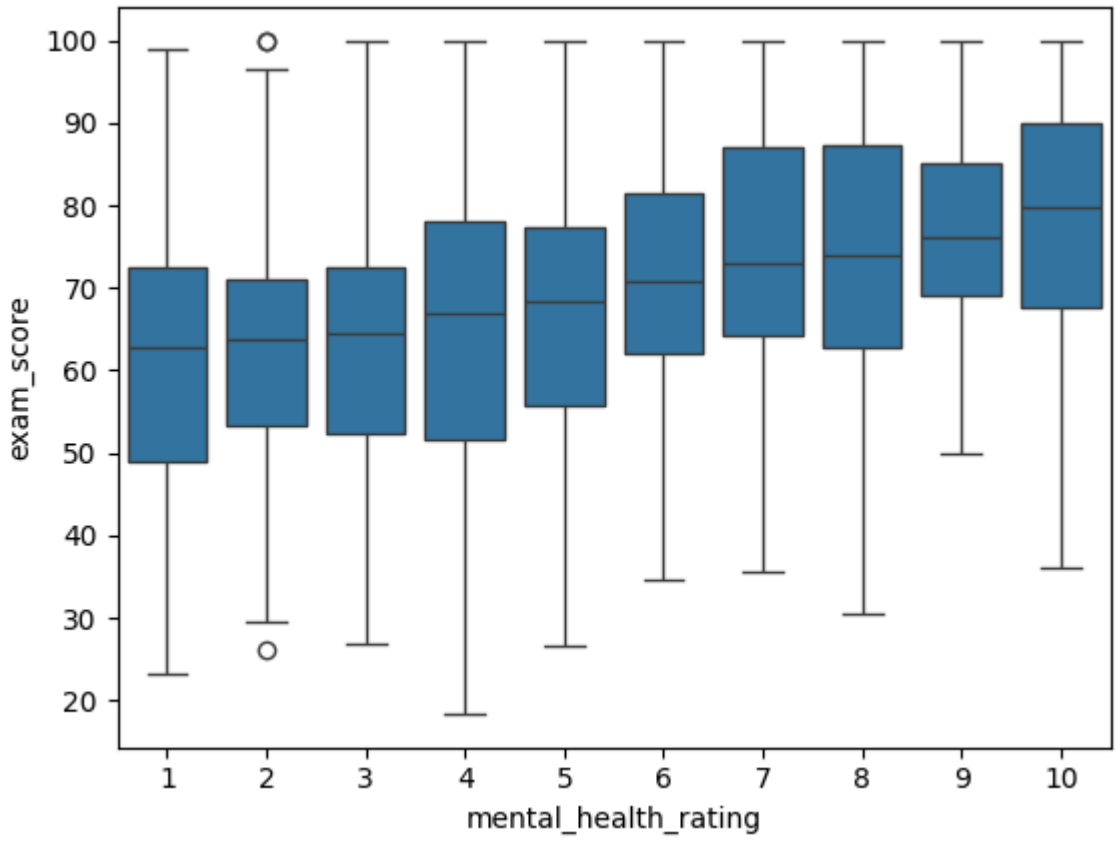
In [9]:
```
#Avaliando notas médias
#por diferentes intervalos (bins) de periodos gastos em redes sociais
# ["0-2h", "2h-4h", "4h-6h", "6h+"]

df["social_media_bin"] = pd.cut(
    df["social_media_hours"],
    bins=[0, 2, 4, 6],
    labels=["0-2h", "2h-4h", "4h-6h"]
)
#Gráfico de caixa
sns.boxplot(x="social_media_bin", y="exam_score", data=df)
plt.title("Notas por tempo em redes sociais")
plt.show()
```

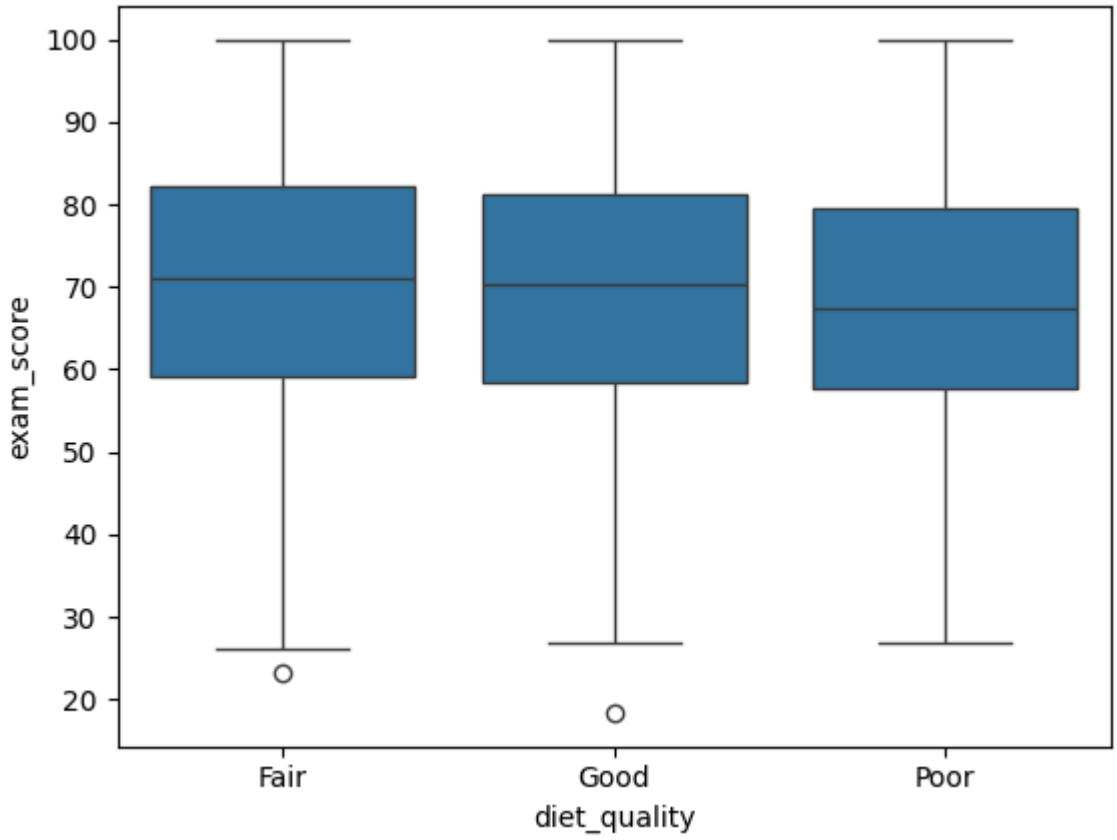


- Alunos mais saudáveis têm melhores desempenhos?

In [10]:
```
#Frequência de exercícios físicos
for col in ["exercise_frequency", "mental_health_rating", "diet_quality"]:
    sns.boxplot(x=col, y="exam_score", data=df)
    plt.show()
```


exam_score
100
90
80
70
60
50
40
30
20
0
1
2
3
4
5
6
exercise_frequency


exam_score
100
90
80
70
60
50
40
30
20
1
2
3
4
5
6
7
8
9
10
mental_health_rating



- Há diferença nas notas entre homens e mulheres?

In [11]:
```
#Estatistica por gênero (média e desvio padrão)
df.groupby(["gender"])["exam_score"].agg(["mean", "std"])
```

Out[11]:

gender	mean	std
Female	69.741372	16.899351
Male	69.368344	17.150875
Other	70.647619	13.755890

In [19]:
```
#Avaliar distribuição de gênero
df["gender"].value_counts(normalize=True)
```

Out[19]:
```
gender
Female    0.481
Male      0.477
Other     0.042
Name: proportion, dtype: float64
```