



```
In [15]: import numpy as np
from datascience import *
import math as m

# These lines do some fancy plotting magic.
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
import warnings
warnings.simplefilter('ignore', FutureWarning)

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import itertools

from ipywidgets import interact, interactive, fixed, interact_manual
import ipywidgets as widgets
```

```
In [9]: #students=Table.read_table("enhanced_student_habits_performance_dataset.csv")
students = pd.read_csv("enhanced_student_habits_performance_dataset.csv")
students
```

```
Out[9]:
```

	student_id	age	gender	major	study_hours_per_day	social_media_
0	100000	26	Male	Computer Science	7.645367	
1	100001	28	Male	Arts	5.700000	
2	100002	17	Male	Arts	2.400000	
3	100003	27	Other	Psychology	3.400000	
4	100004	25	Female	Business	4.700000	
...
79995	179995	16	Male	Engineering	3.700000	
79996	179996	16	Female	Business	1.200000	
79997	179997	26	Female	Arts	4.100000	
79998	179998	23	Other	Biology	3.800000	
79999	179999	25	Other	Engineering	3.000000	

80000 rows × 31 columns

```
In [12]: def plot_variable_associations(df, max_plots=3000):
    """
    Iterates through pairs of variables in df and plots them based on type.
    Limits number of plots with max_plots for readability.
```

```
"""
num_cols = df.select_dtypes(include=['int64', 'float64']).columns
cat_cols = df.select_dtypes(include=['object', 'category', 'bool']).columns

all_vars = list(num_cols) + list(cat_cols)
combos = list(itertools.combinations(all_vars, 2))

for i, (var1, var2) in enumerate(combos):
    if i >= max_plots:
        print(f"Stopped after {max_plots} plots for readability.")
        break

plt.figure(figsize=(6,4))

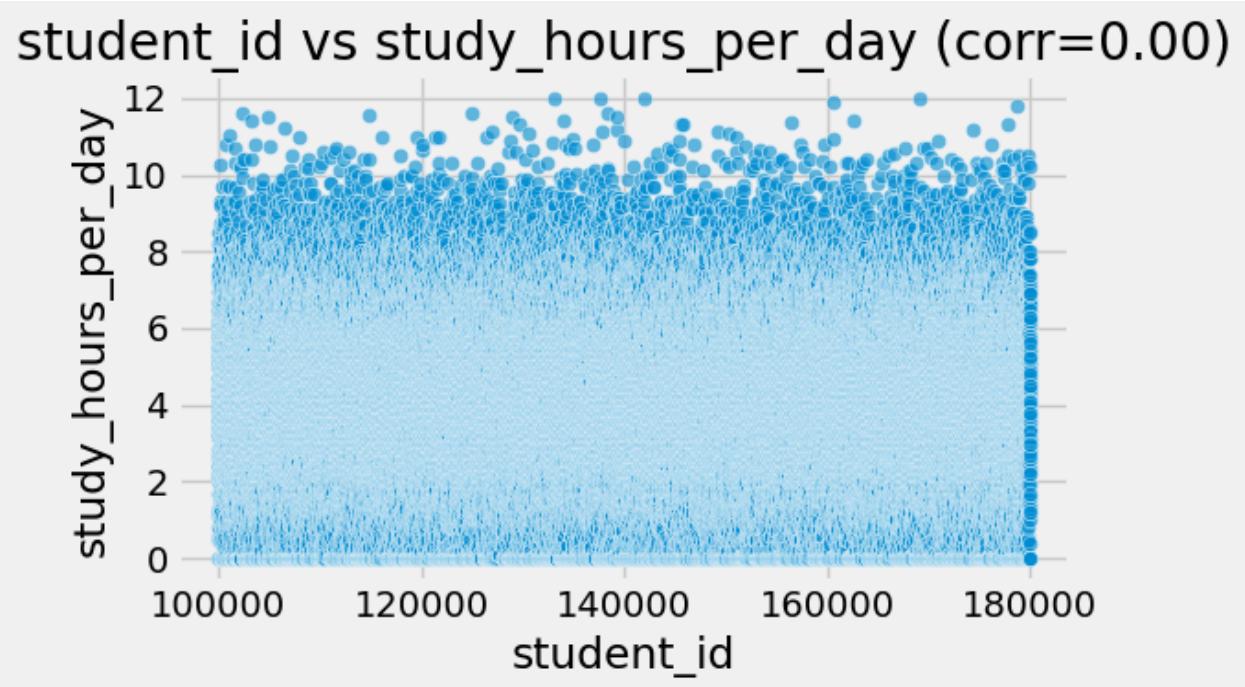
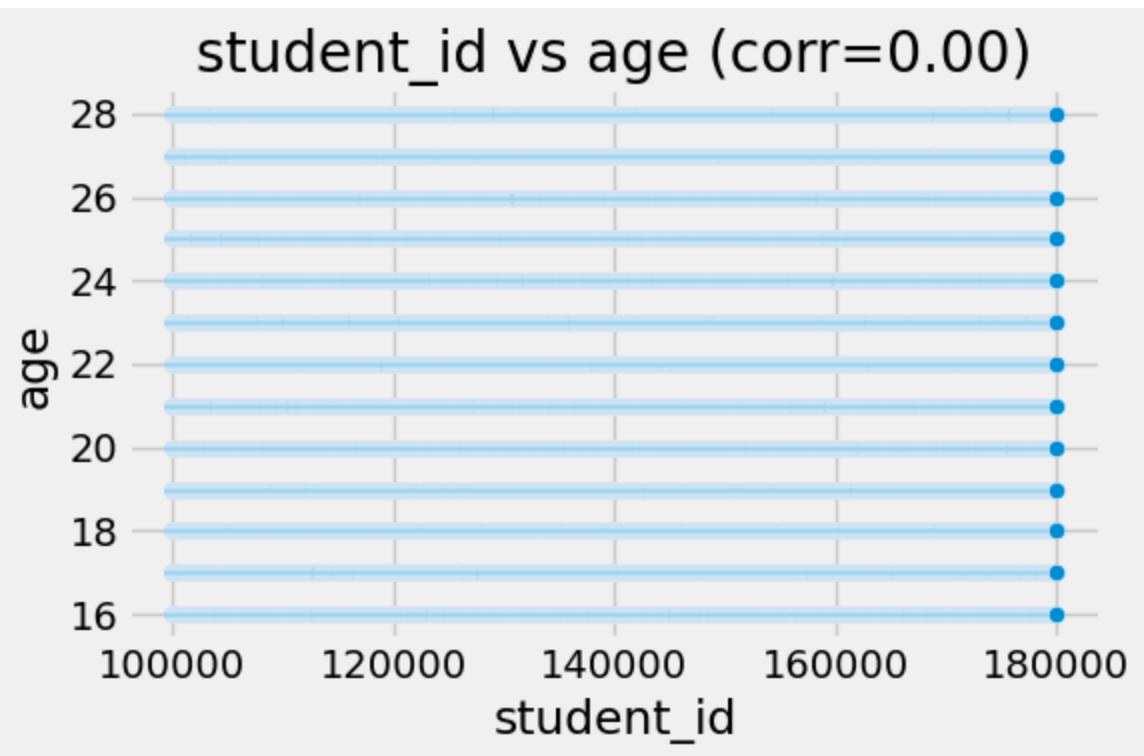
# Numerical vs Numerical
if var1 in num_cols and var2 in num_cols:
    sns.scatterplot(data=df, x=var1, y=var2, alpha=0.6)
    corr = df[[var1, var2]].corr().iloc[0,1]
    plt.title(f"{var1} vs {var2} (corr={corr:.2f})")

# Numerical vs Categorical
elif var1 in cat_cols and var2 in num_cols:
    sns.boxplot(data=df, x=var1, y=var2)
    plt.title(f"{var2} by {var1}")
elif var2 in cat_cols and var1 in num_cols:
    sns.boxplot(data=df, x=var2, y=var1)
    plt.title(f"{var1} by {var2}")

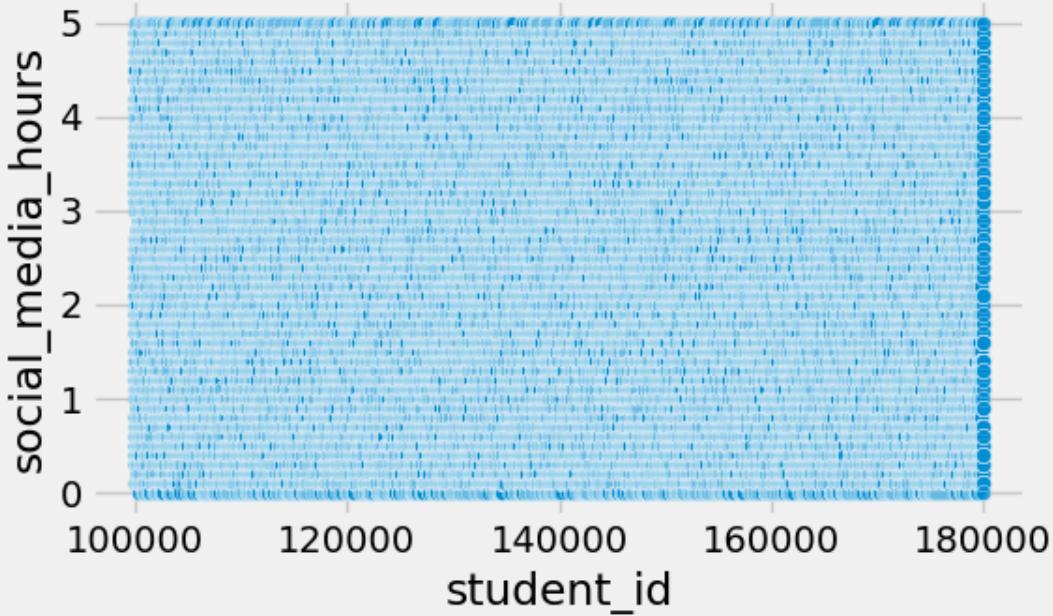
# Categorical vs Categorical
else:
    ctab = pd.crosstab(df[var1], df[var2])
    sns.heatmap(ctab, annot=True, fmt="d", cmap="Blues")
    plt.title(f"{var1} vs {var2} (counts)")

plt.tight_layout()
plt.show()

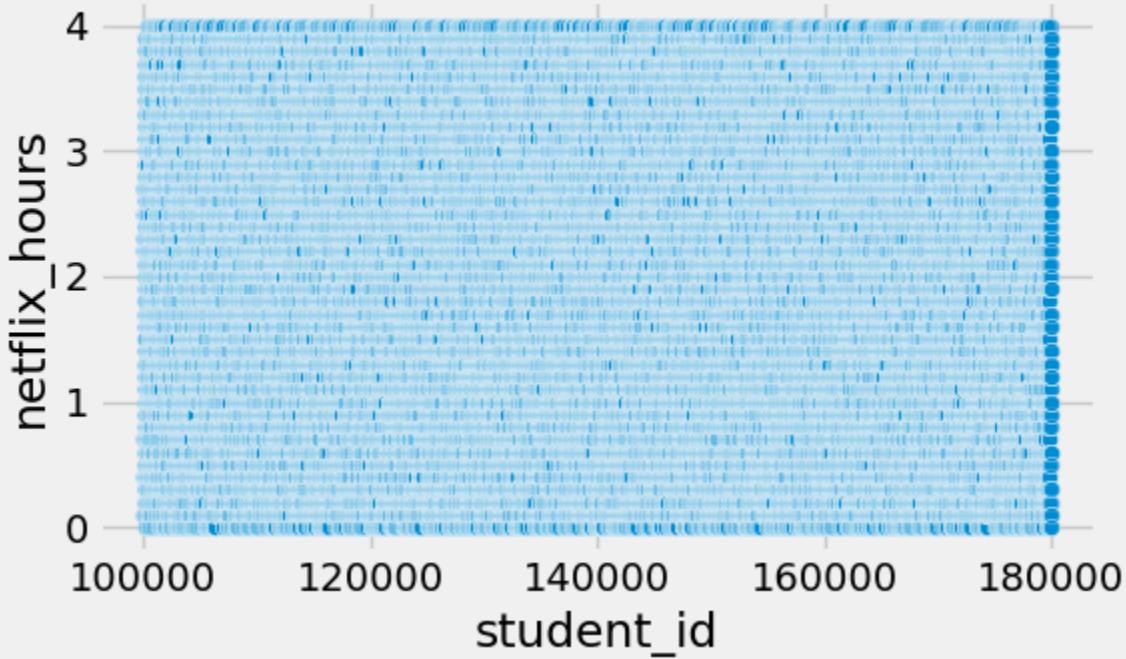
plot_variable_associations(students)
```



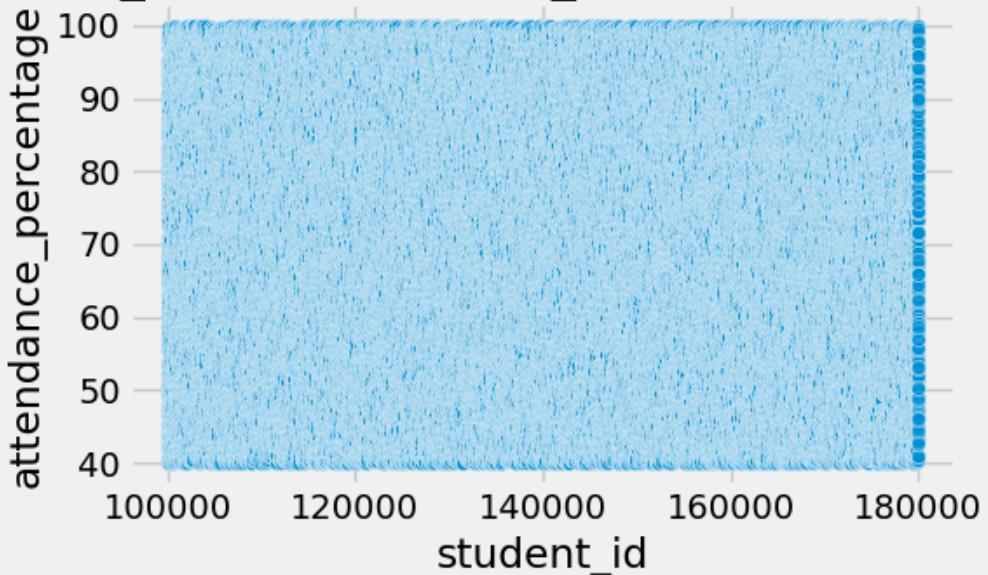
student_id vs social_media_hours (corr=-0.00)



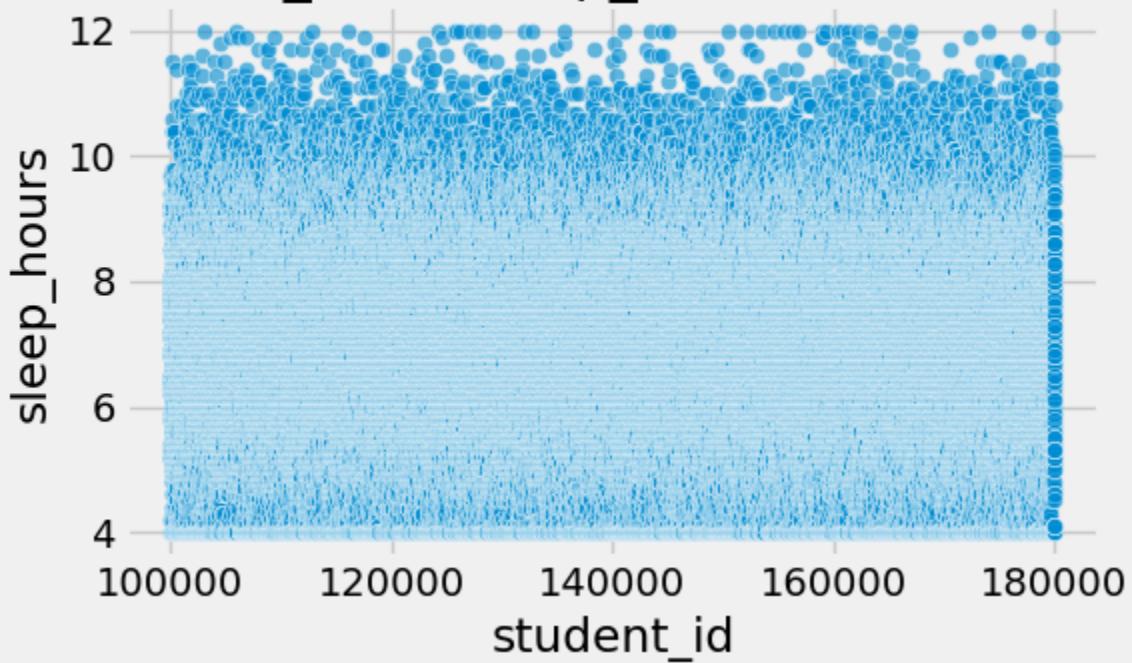
student_id vs netflix_hours (corr=-0.01)



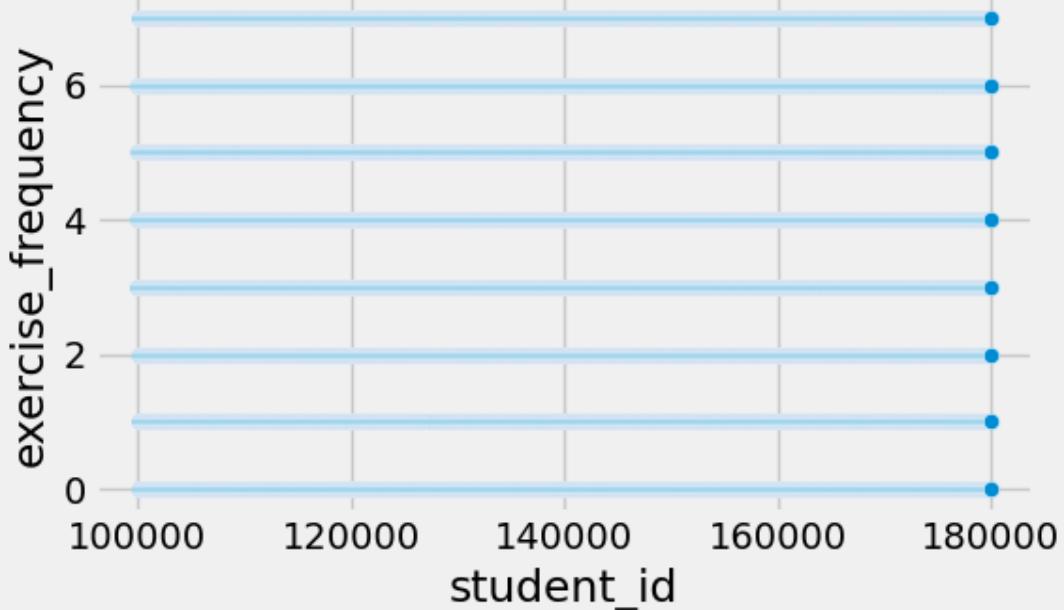
student_id vs attendance_percentage (corr=0.00)



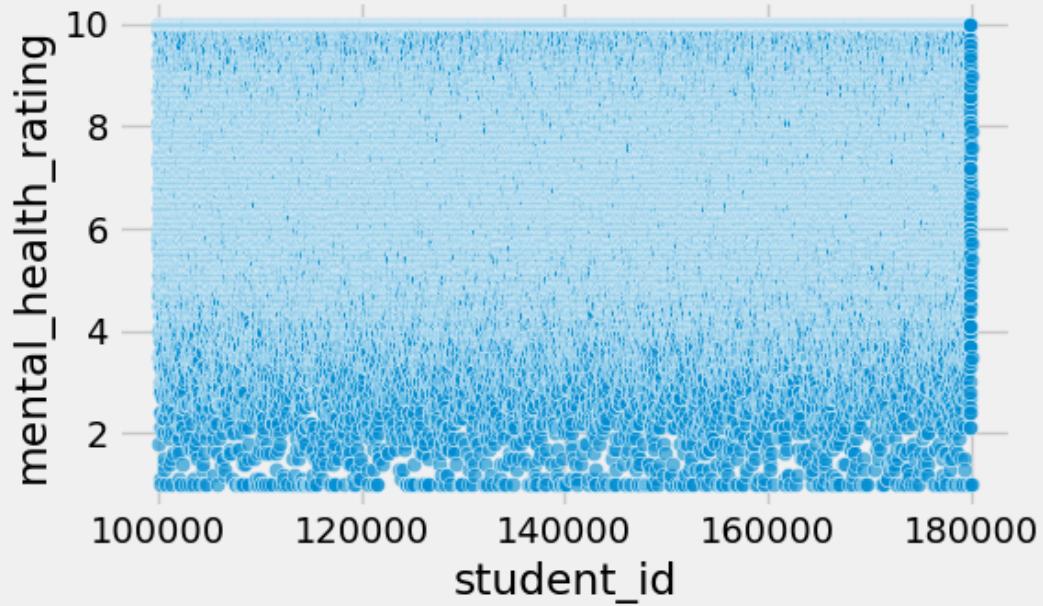
student_id vs sleep_hours (corr=-0.00)



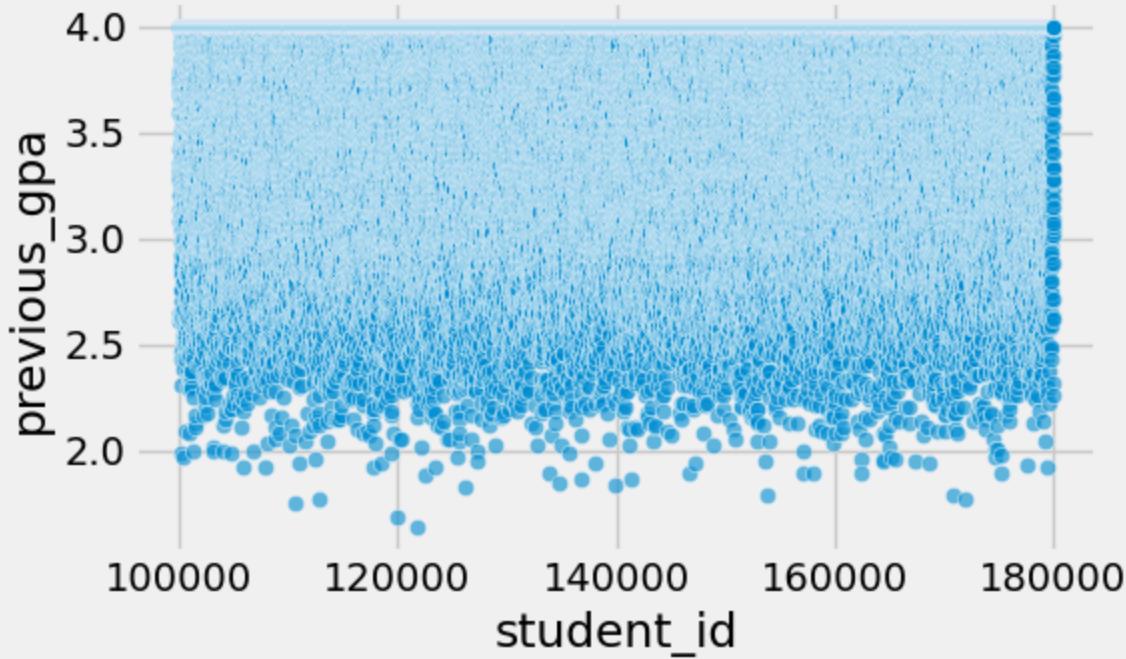
student_id vs exercise_frequency (corr=-0.00)



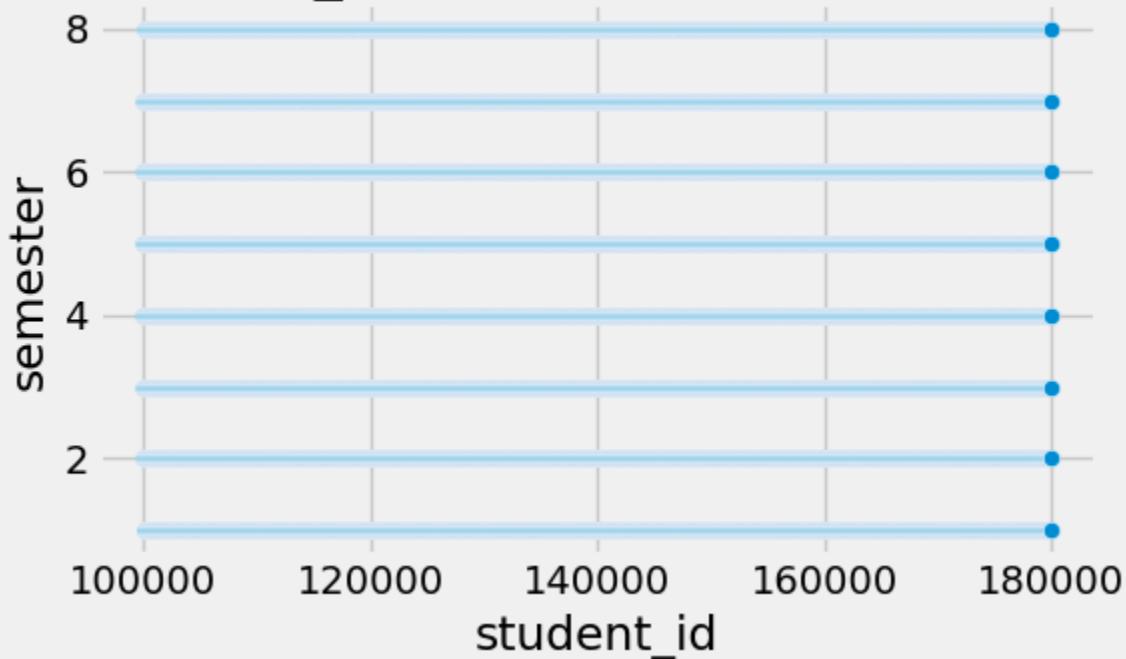
student_id vs mental_health_rating (corr=0.00)



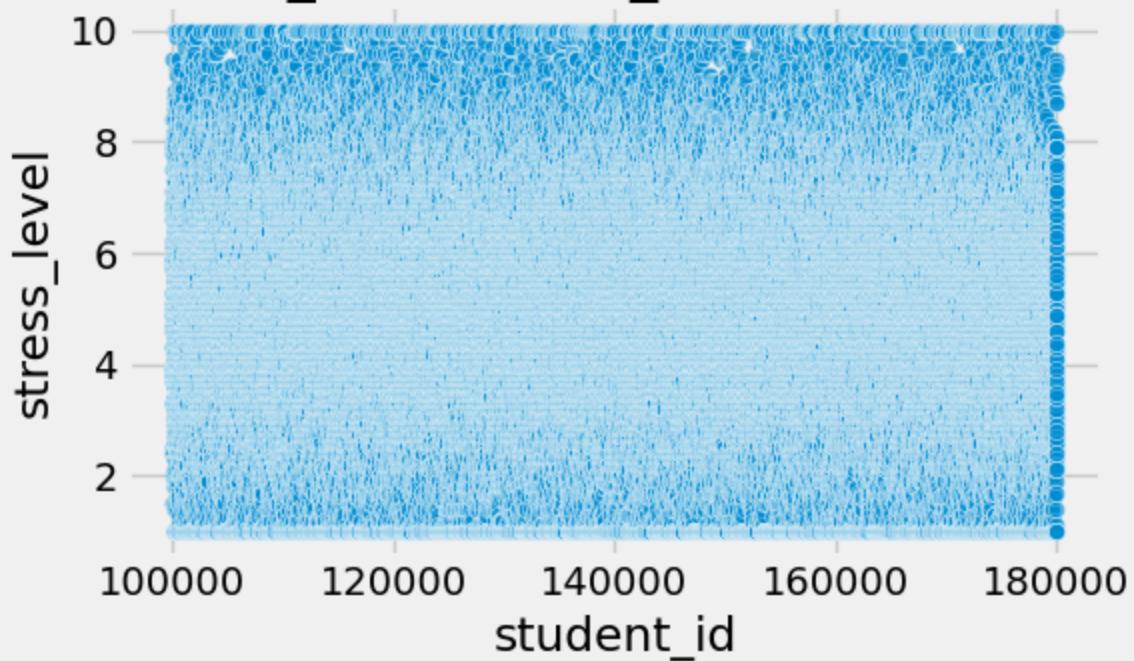
student_id vs previous_gpa (corr=0.01)



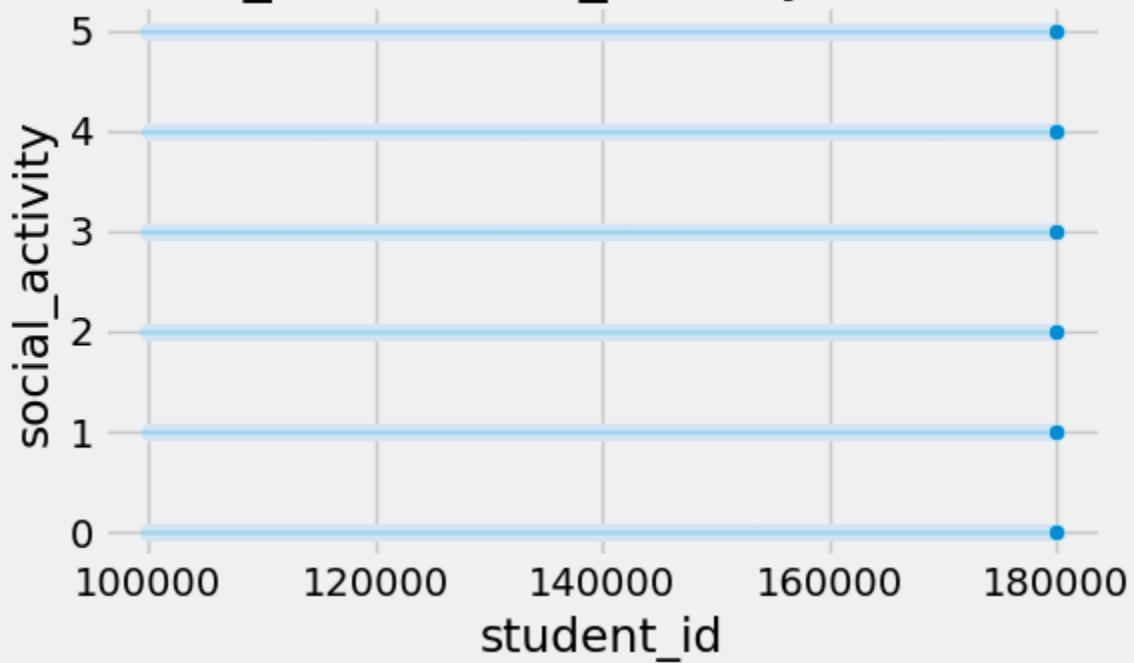
student_id vs semester (corr=0.00)



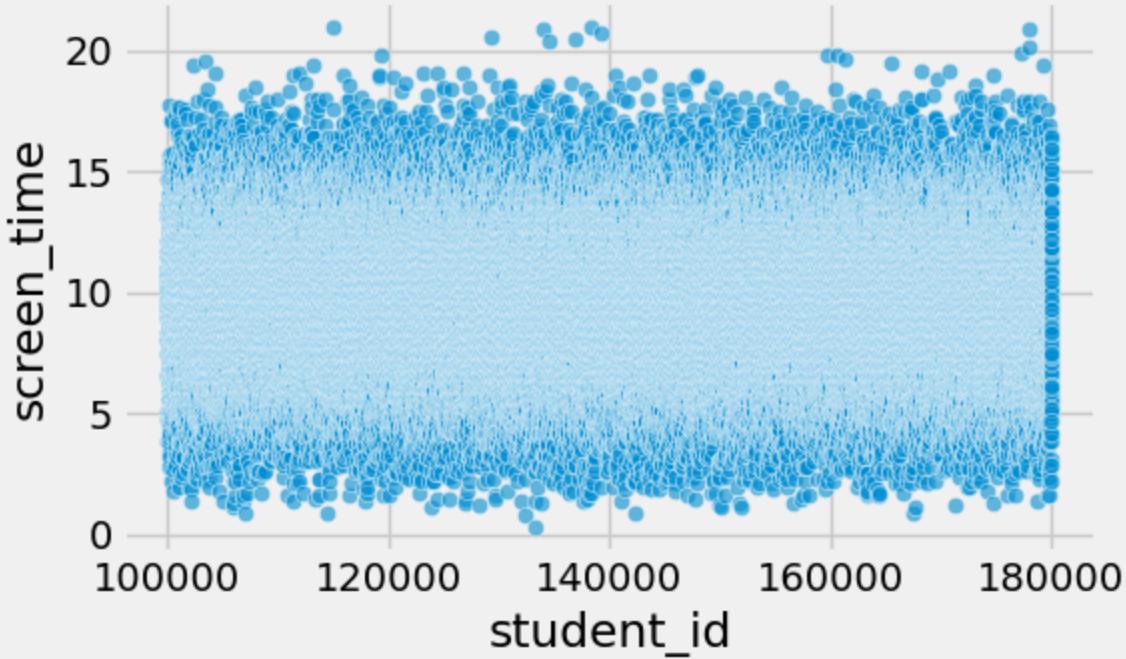
student_id vs stress_level (corr=-0.00)



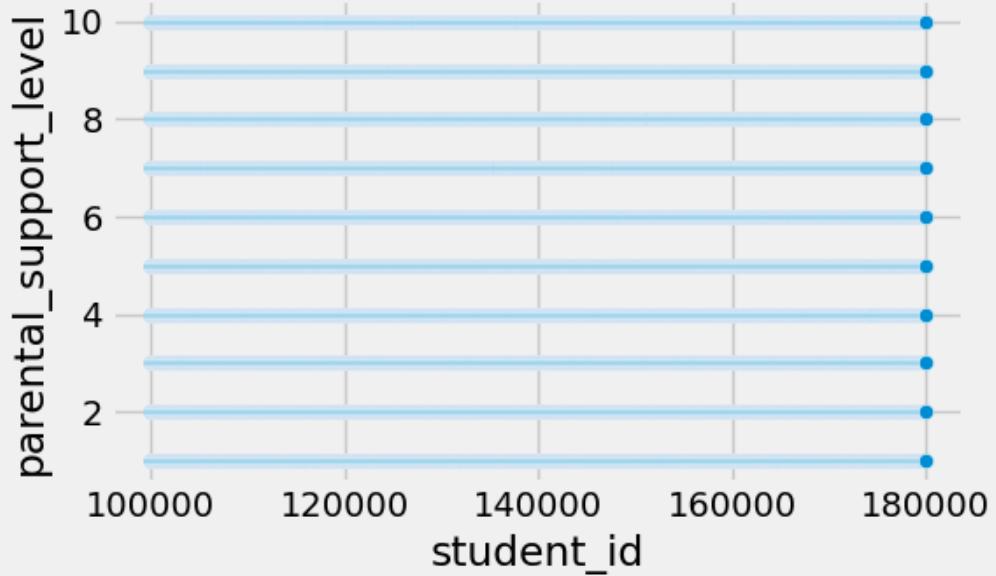
student_id vs social_activity (corr=-0.01)



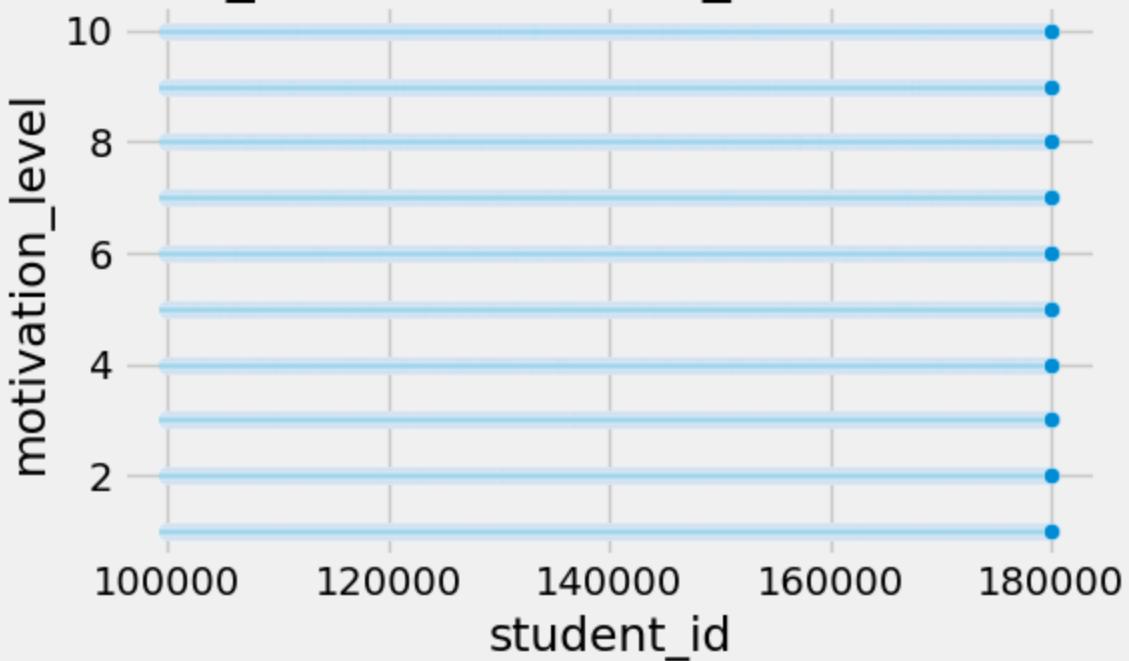
student_id vs screen_time (corr=-0.00)



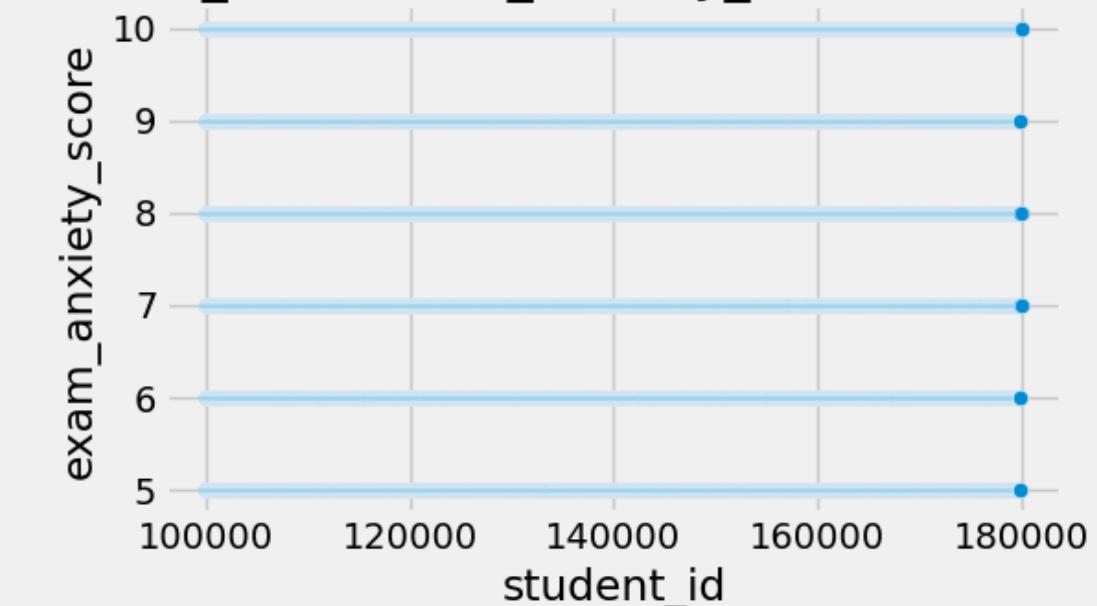
student_id vs parental_support_level (corr=-0.00)



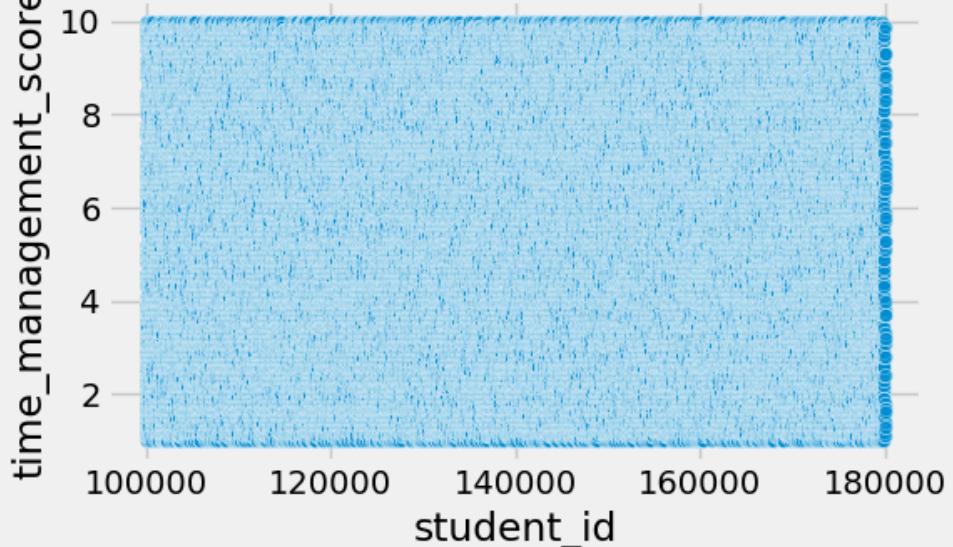
student_id vs motivation_level (corr=0.00)



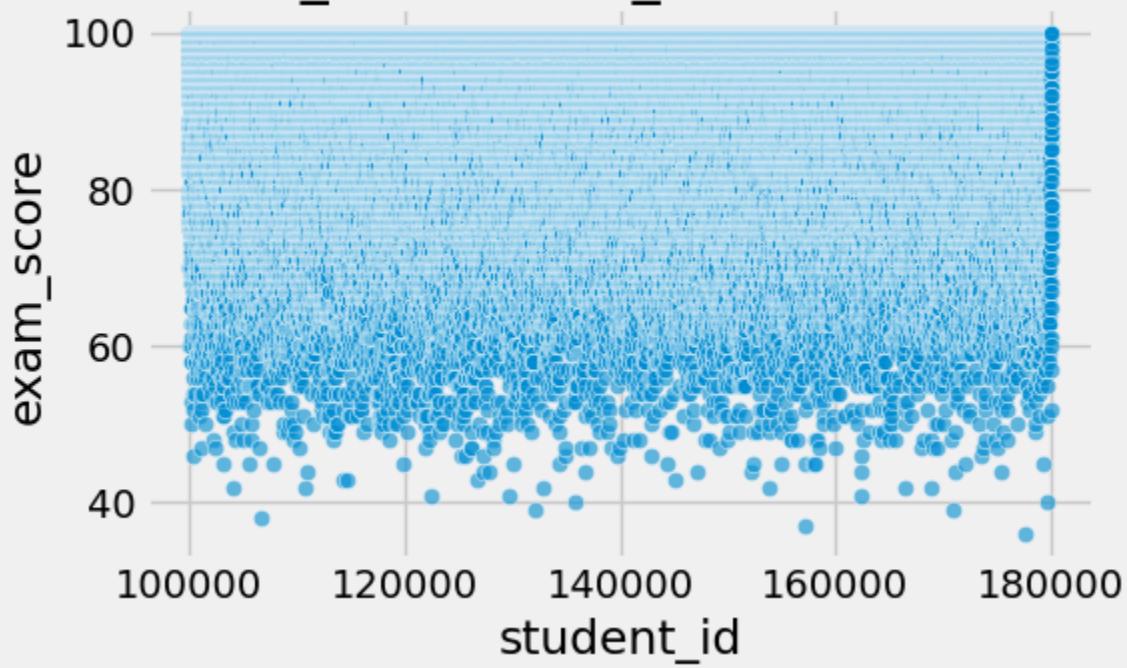
student_id vs exam_anxiety_score (corr=-0.00)

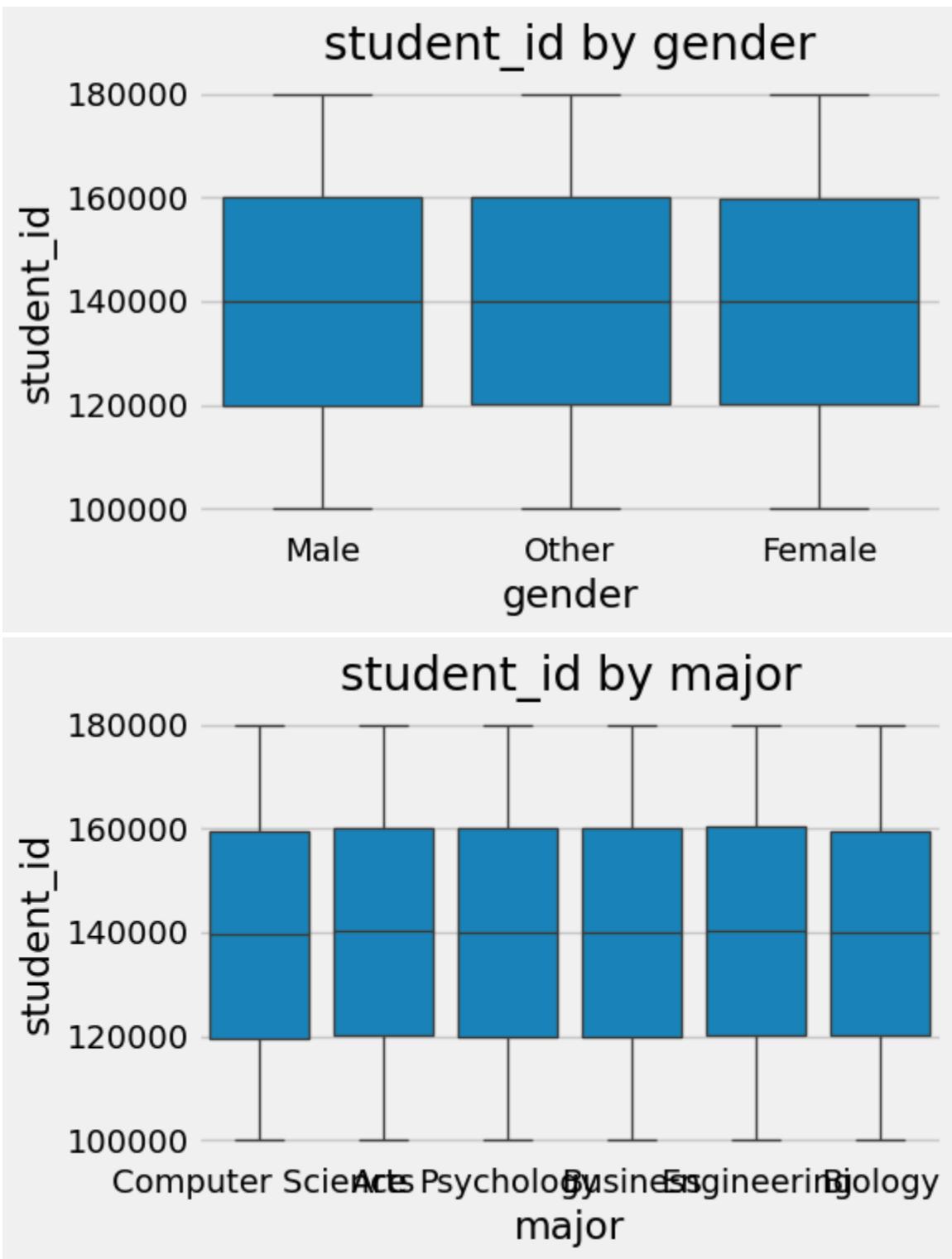


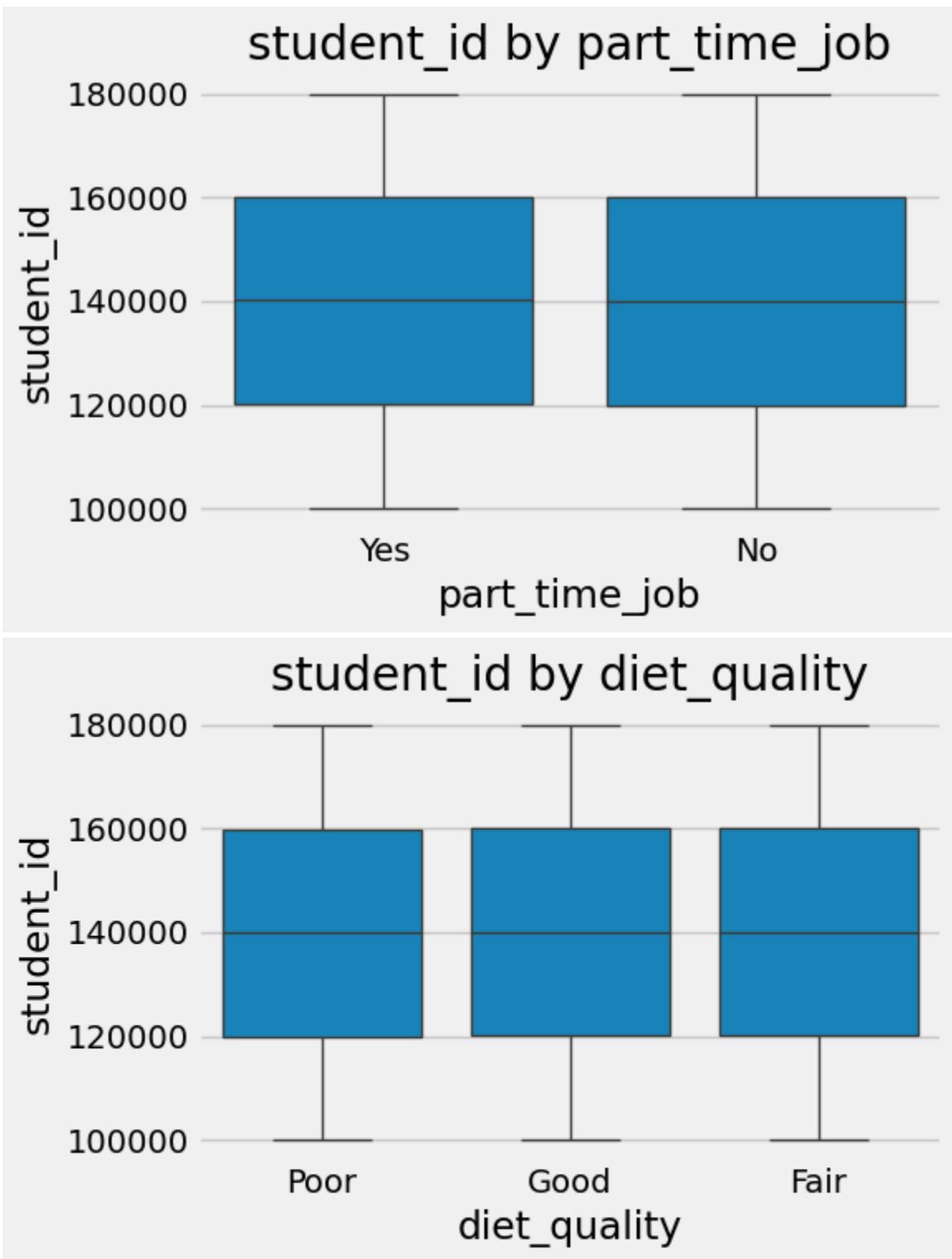
student_id vs time_management_score (corr=-0.00)

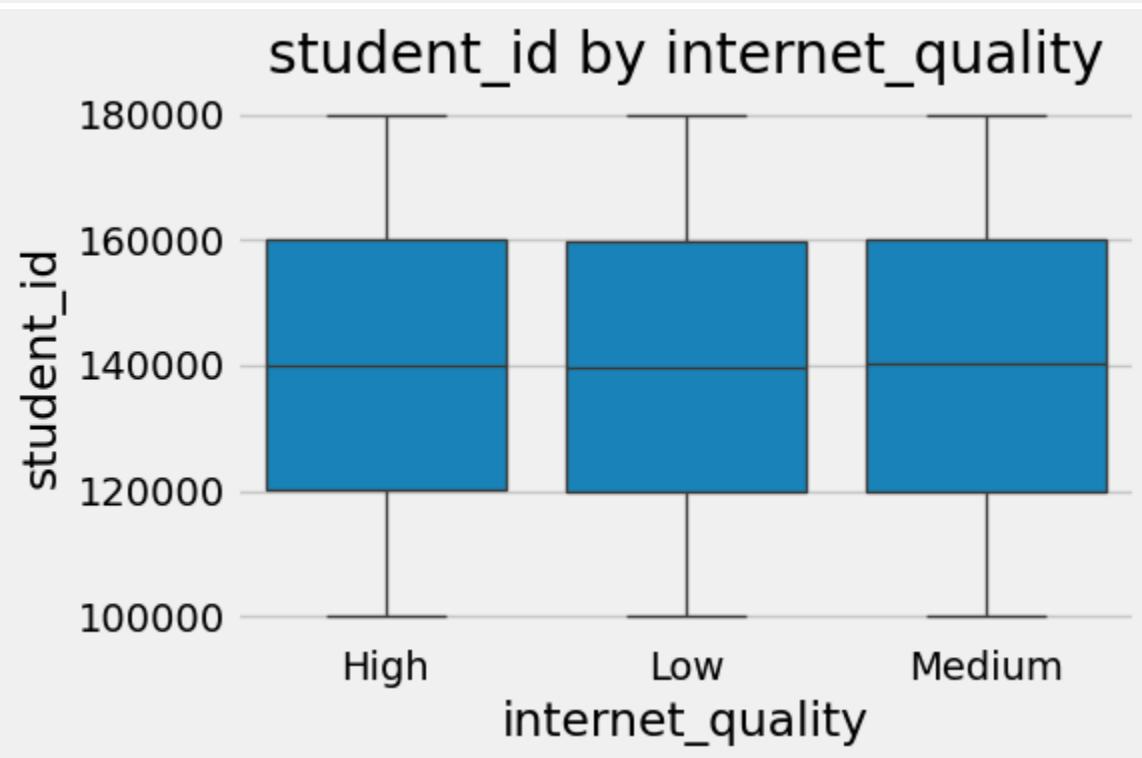
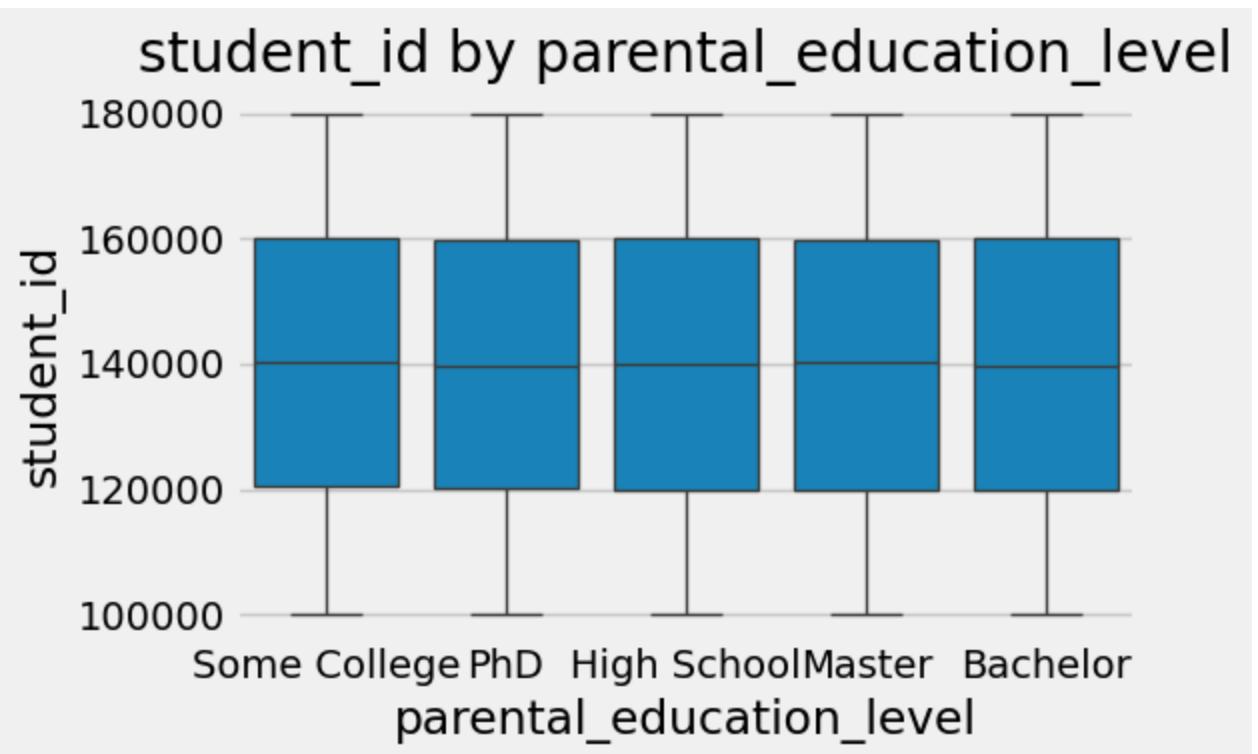


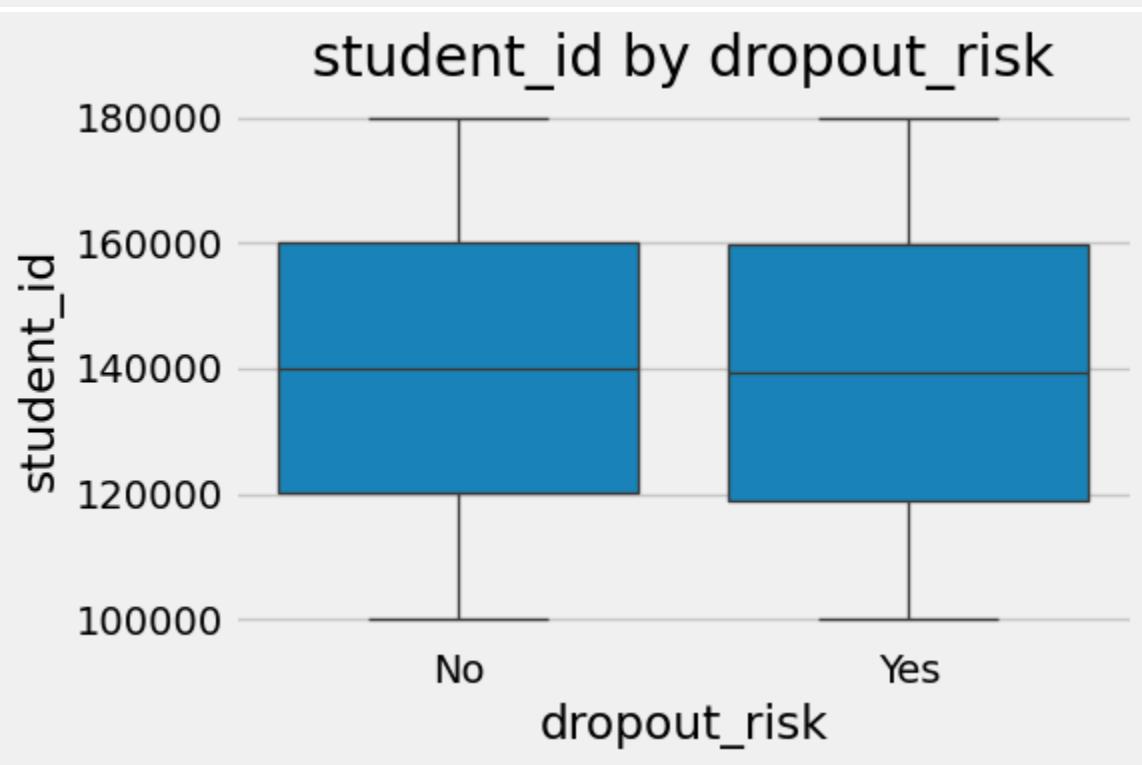
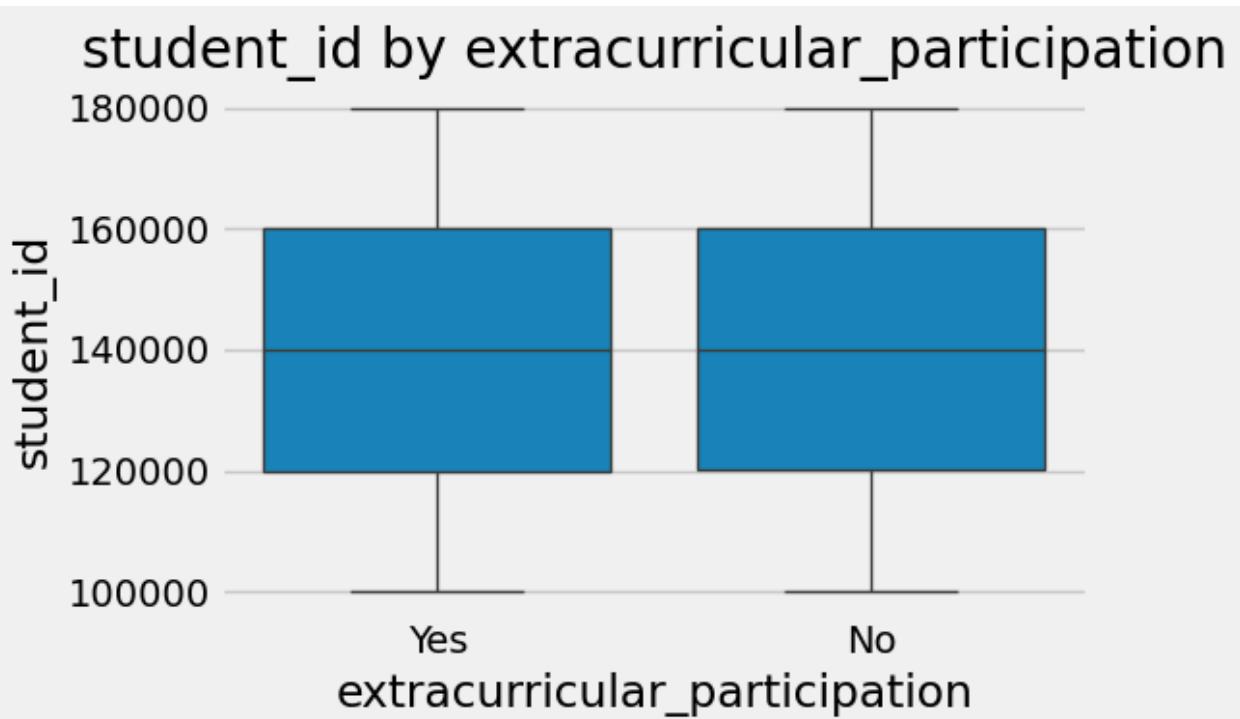
student_id vs exam_score (corr=0.01)

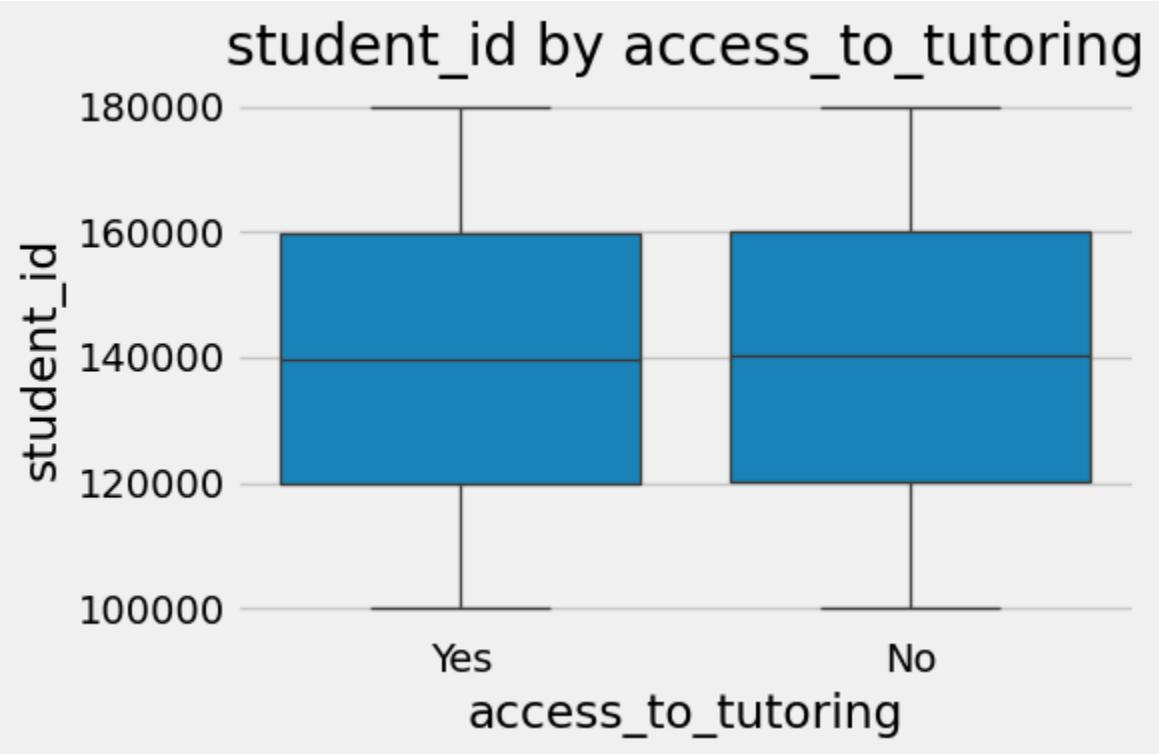
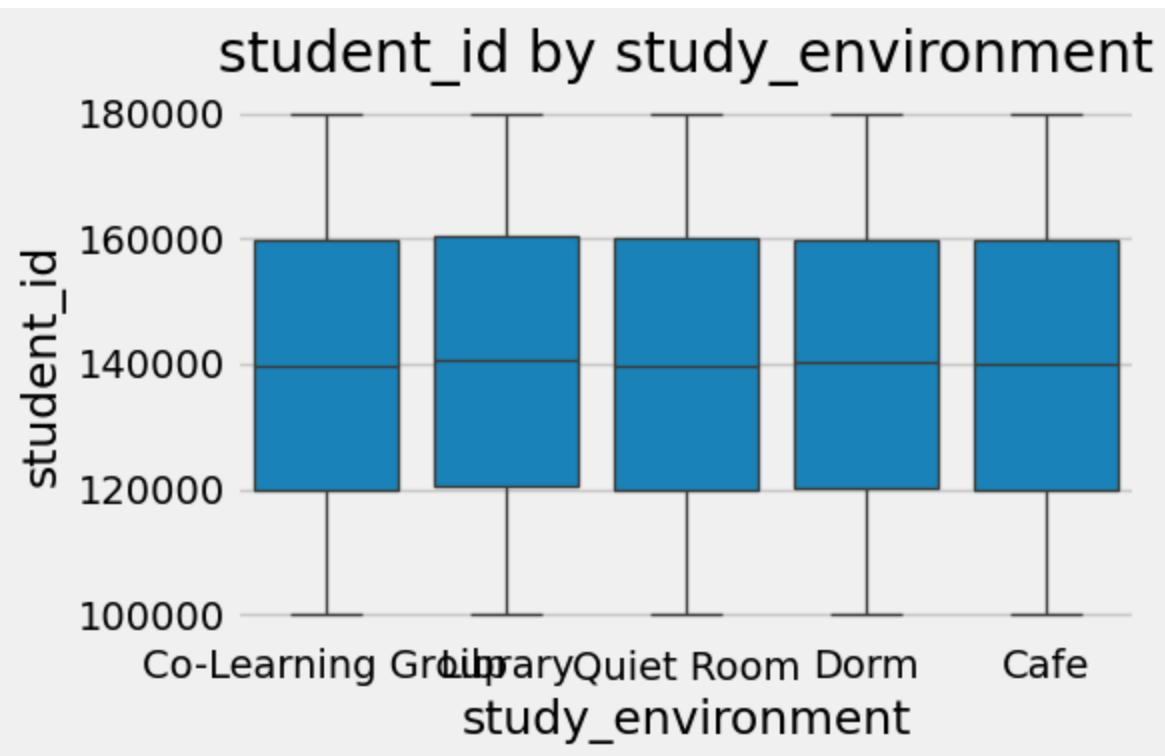


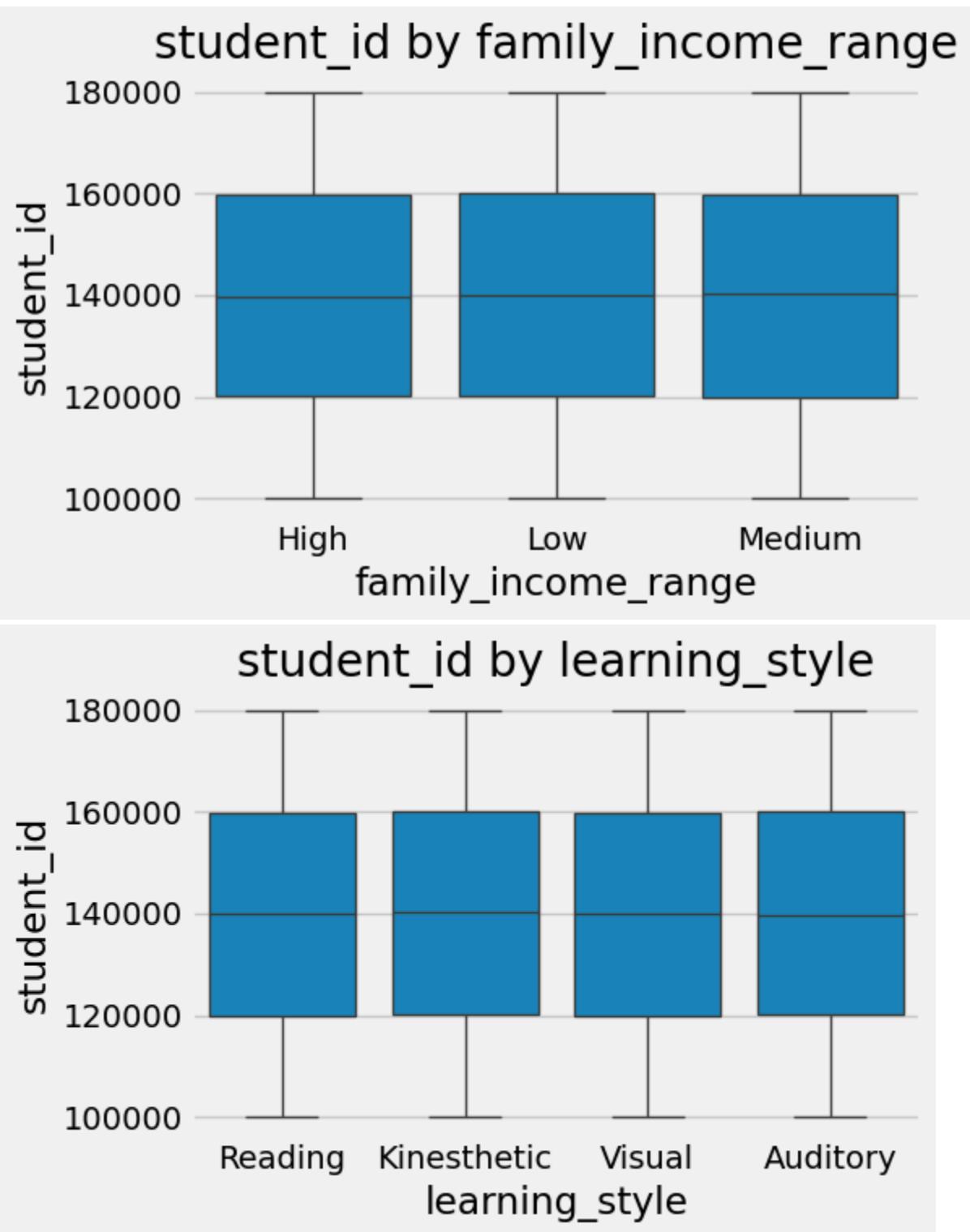




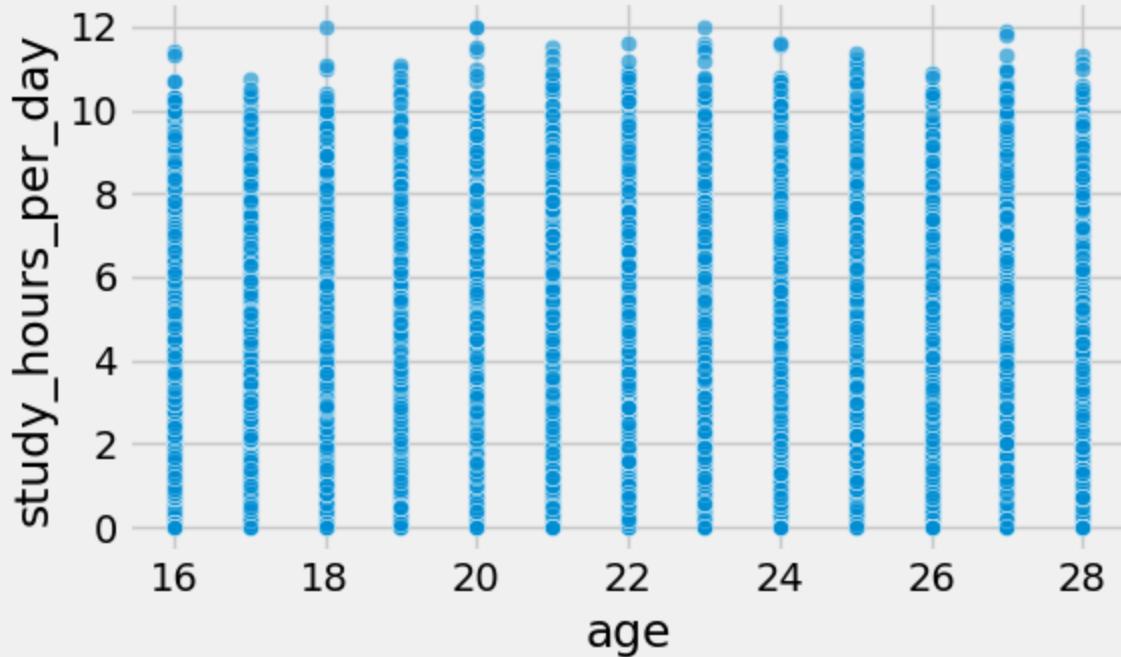




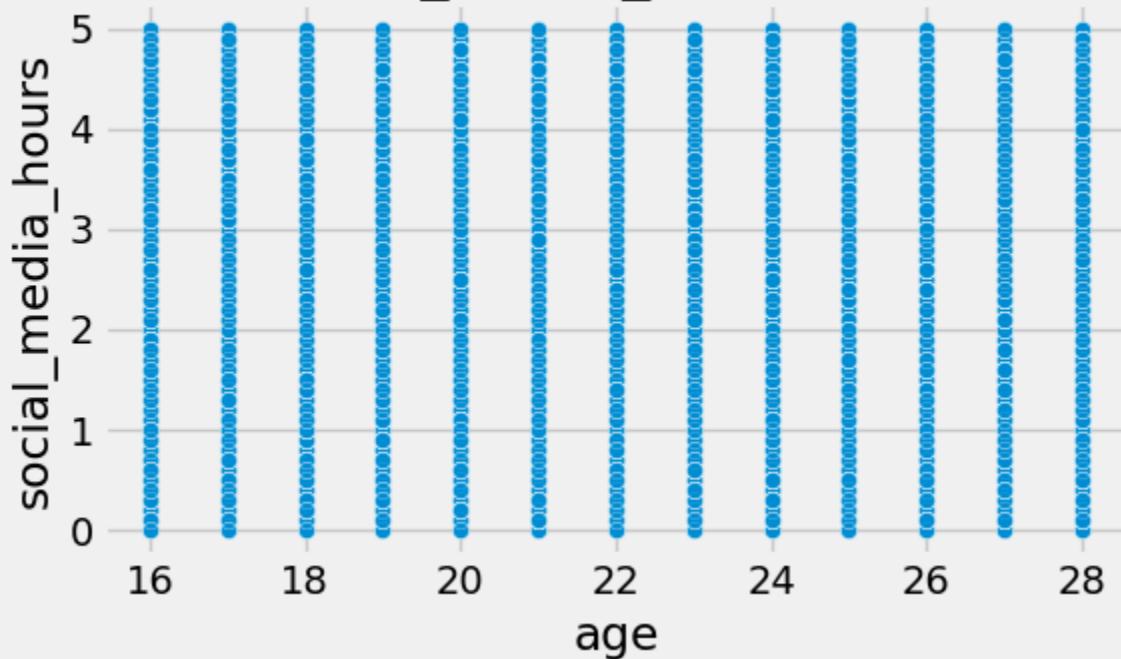




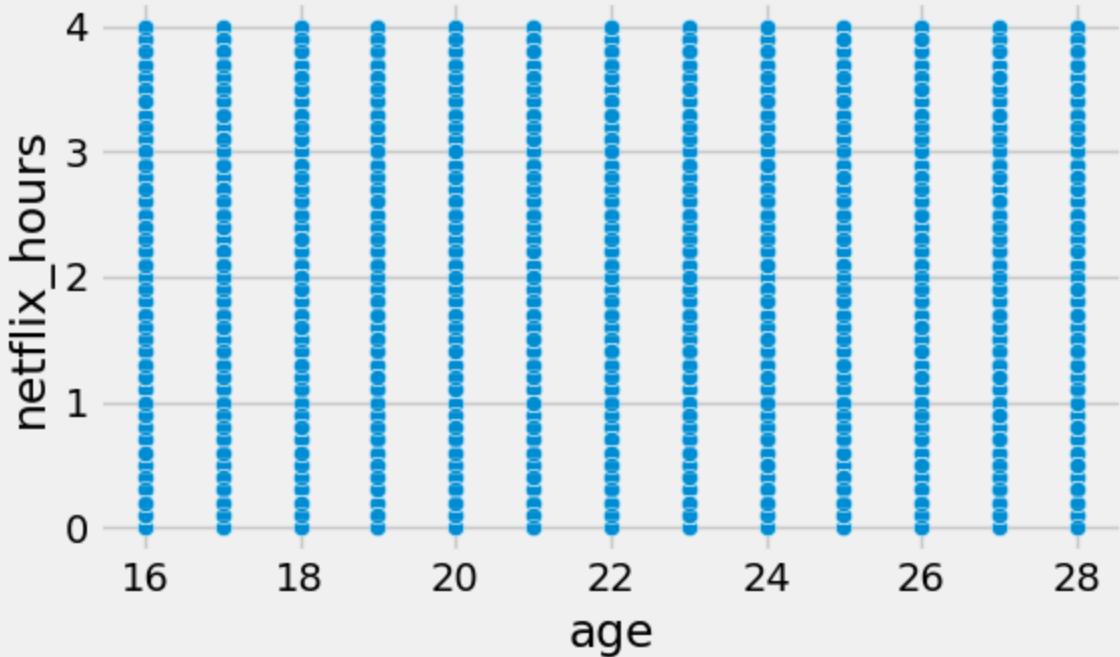
age vs study_hours_per_day (corr=-0.00)



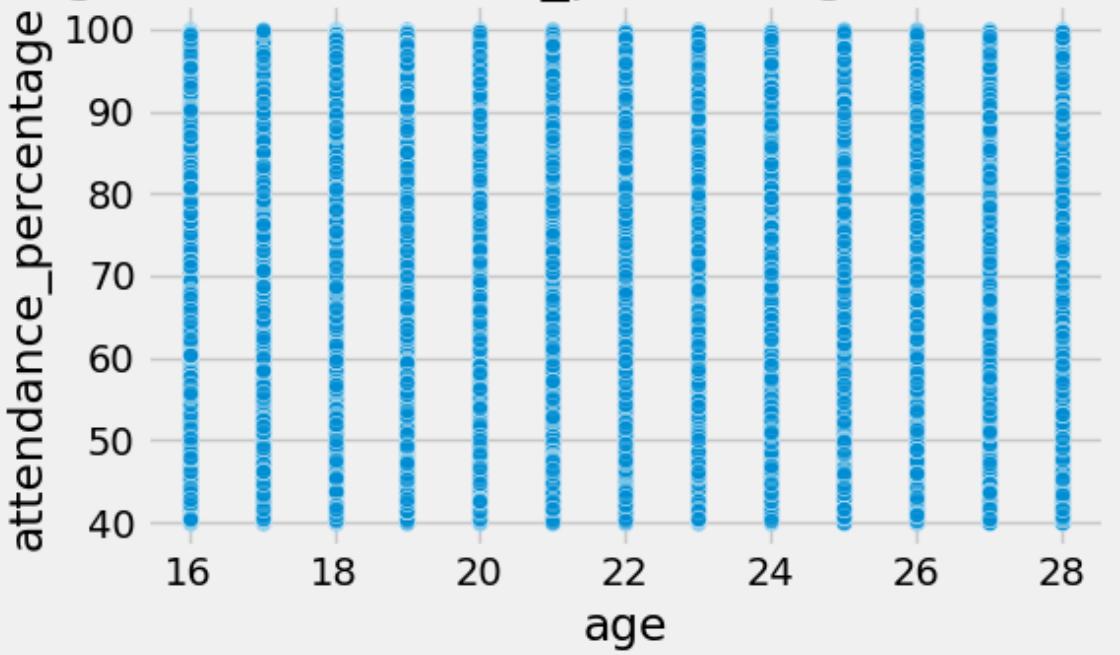
age vs social_media_hours (corr=-0.01)

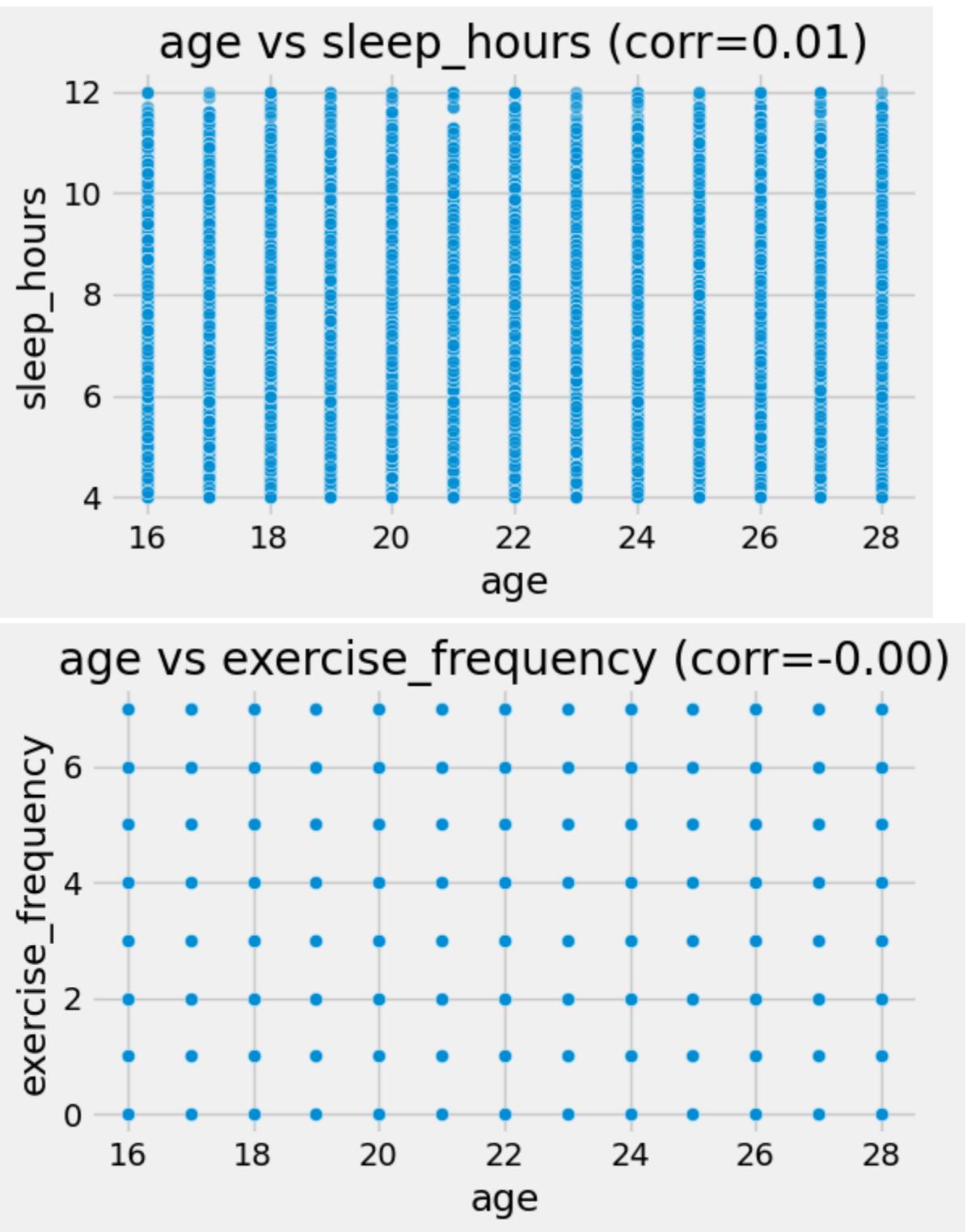


age vs netflix_hours (corr=-0.00)

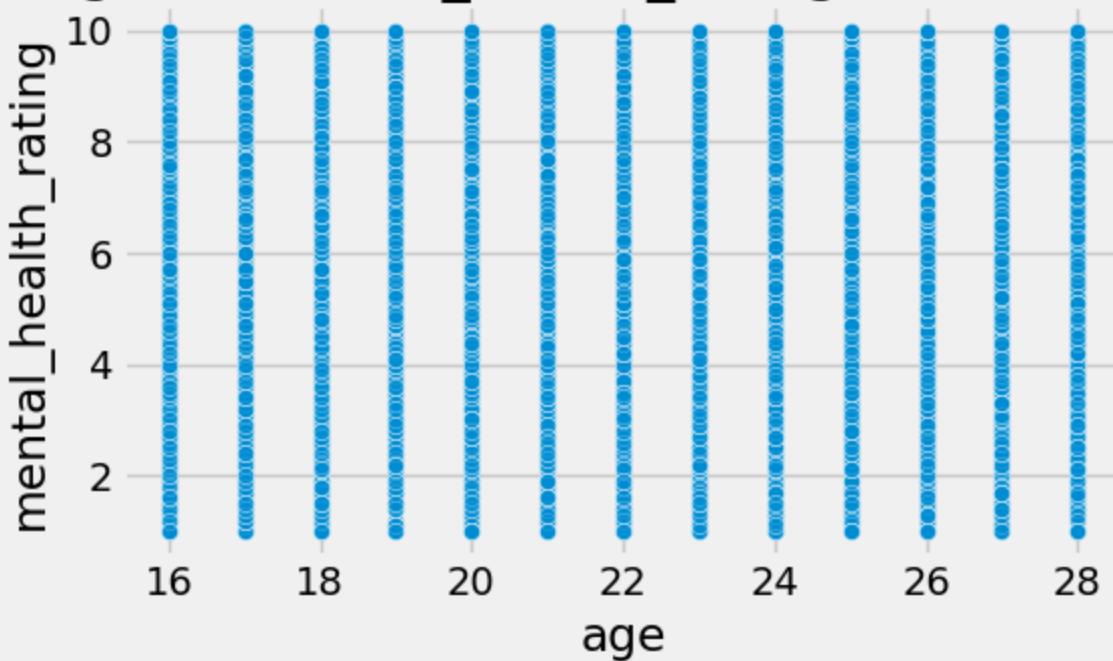


age vs attendance_percentage (corr=0.00)

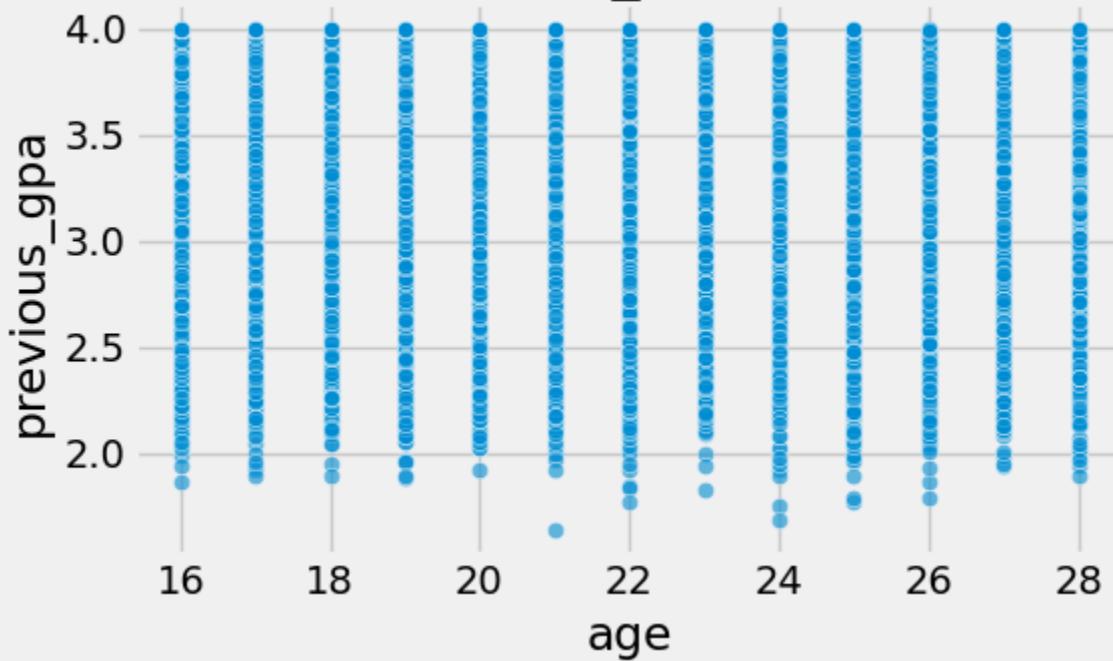




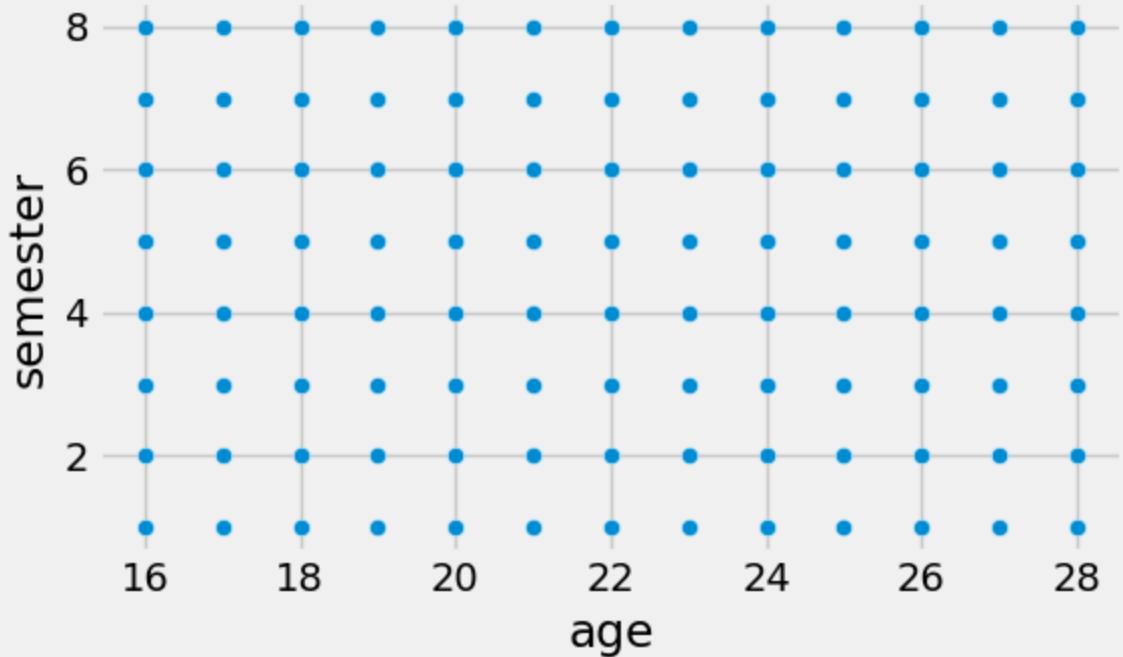
age vs mental_health_rating (corr=-0.00)



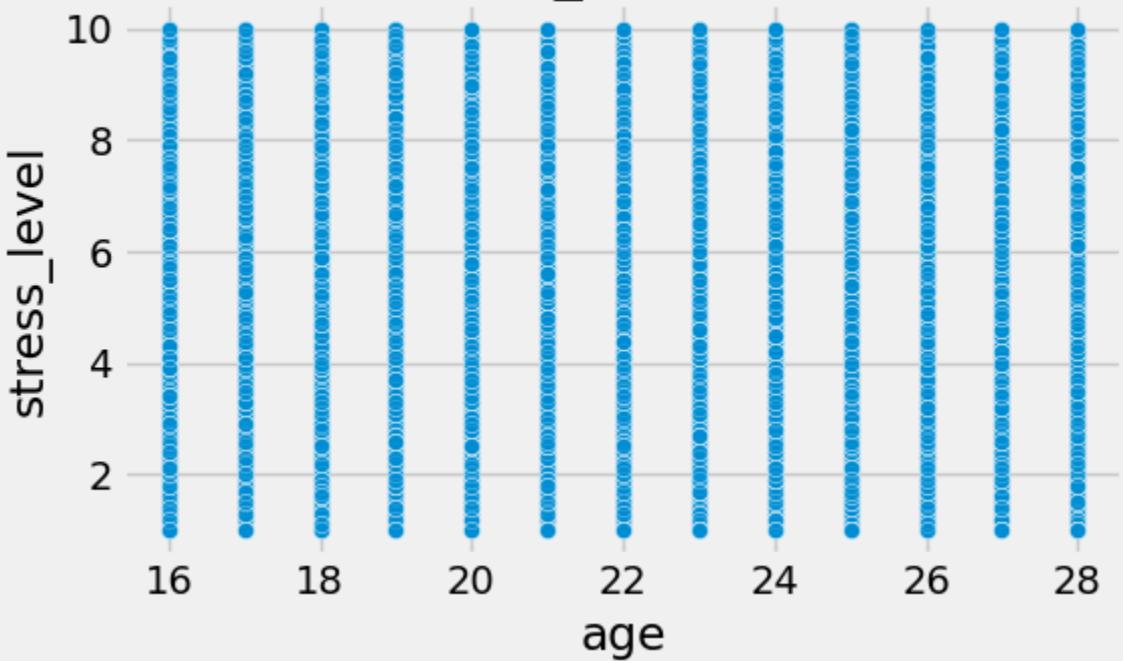
age vs previous_gpa (corr=0.00)

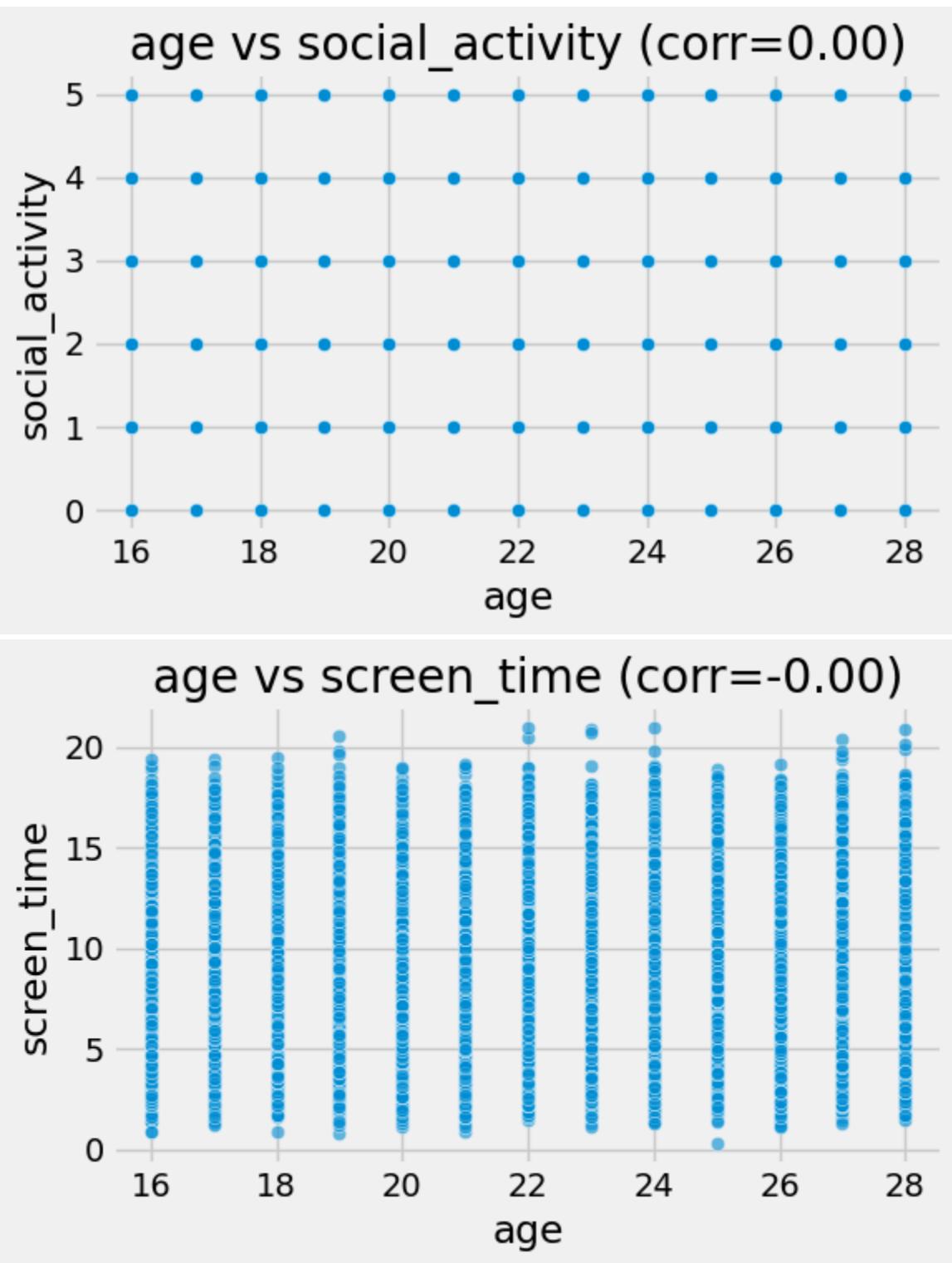


age vs semester (corr=-0.00)

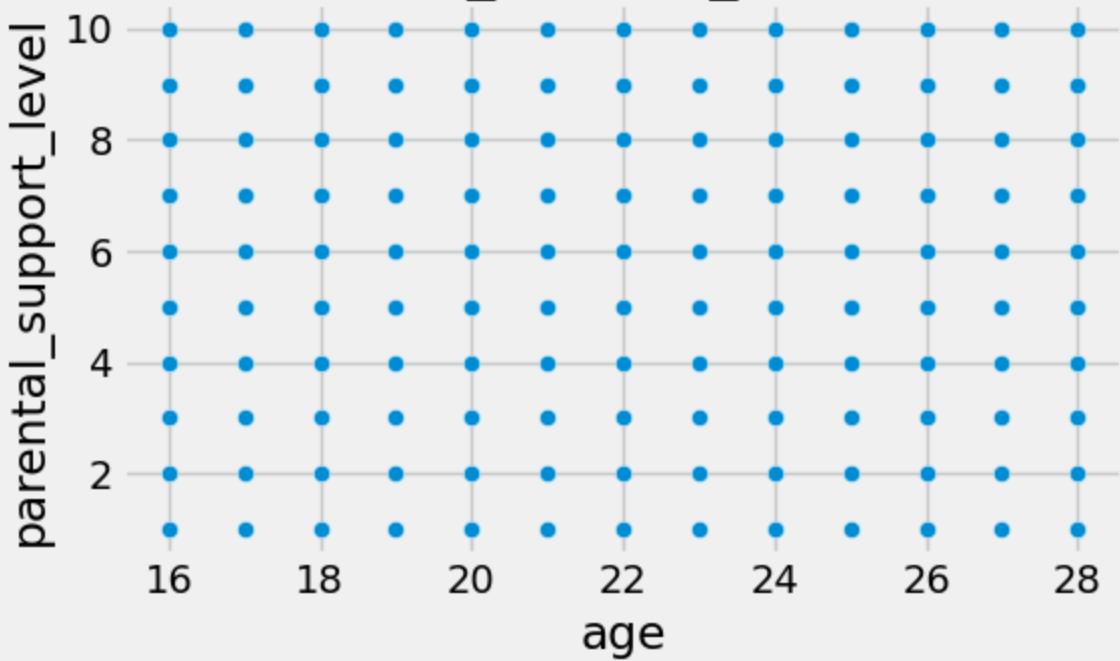


age vs stress_level (corr=0.00)

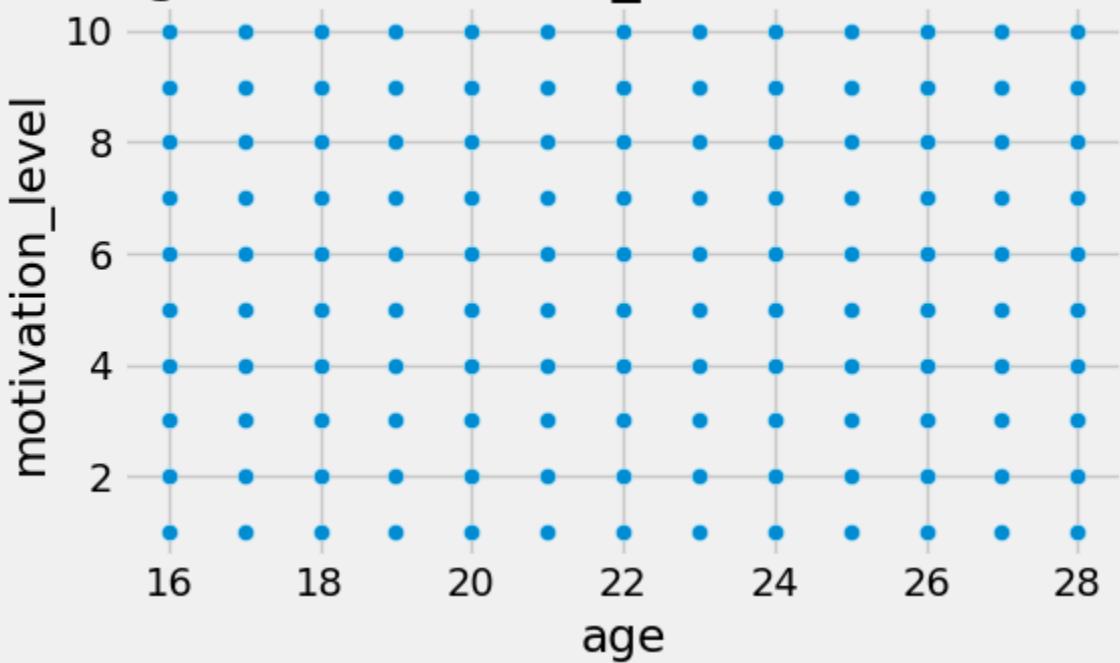




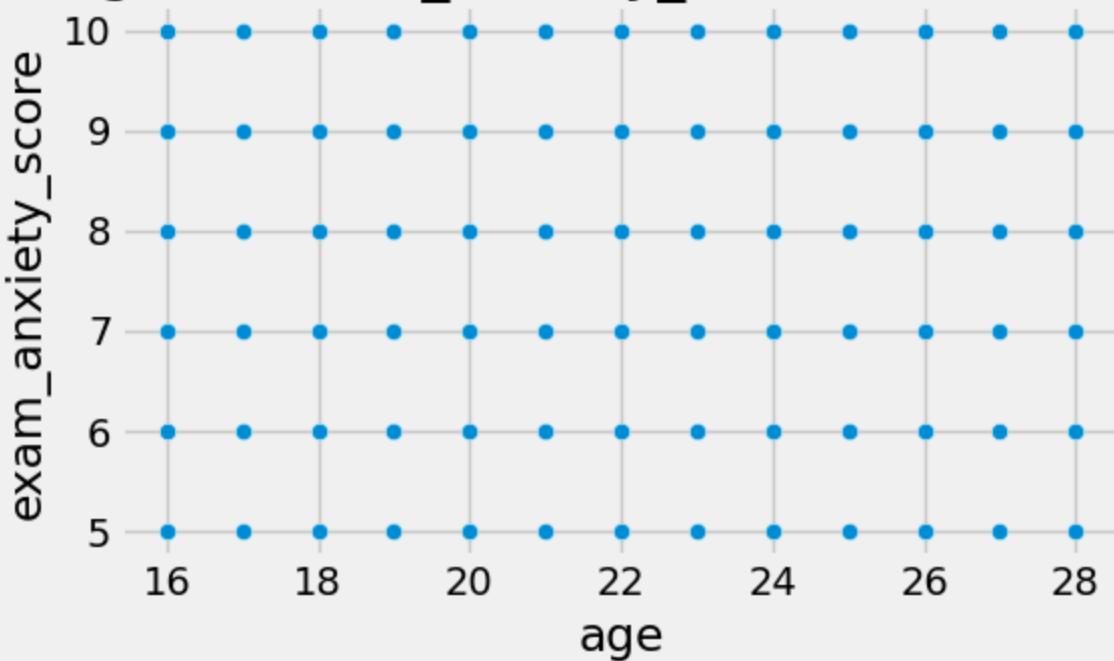
age vs parental_support_level (corr=0.00)



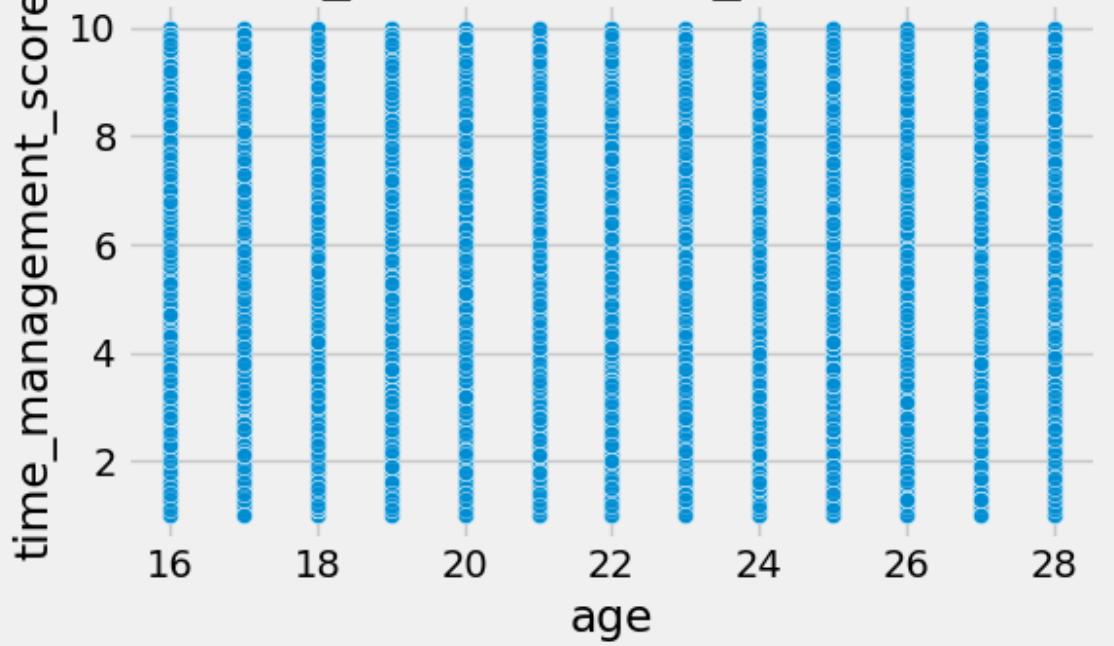
age vs motivation_level (corr=-0.00)

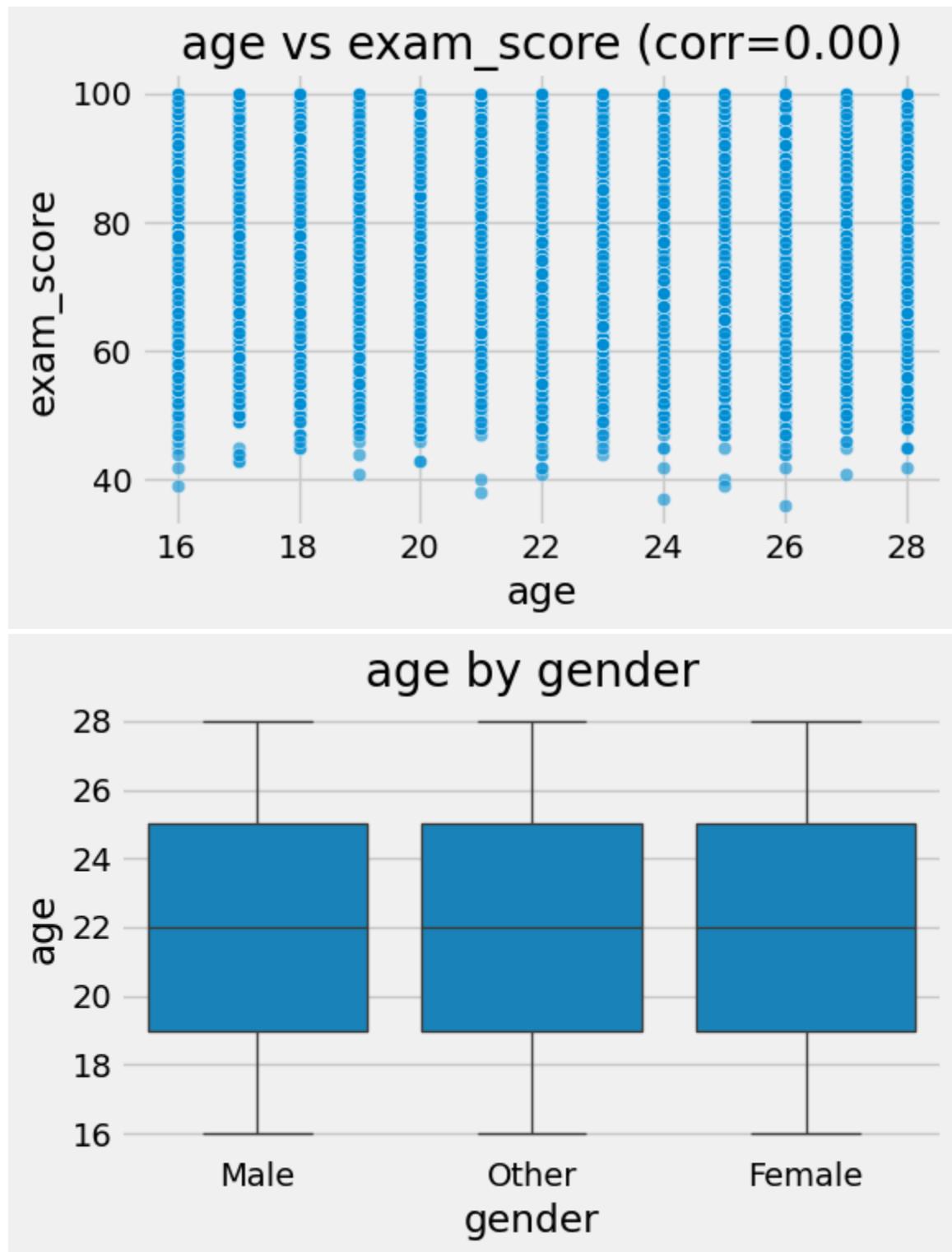


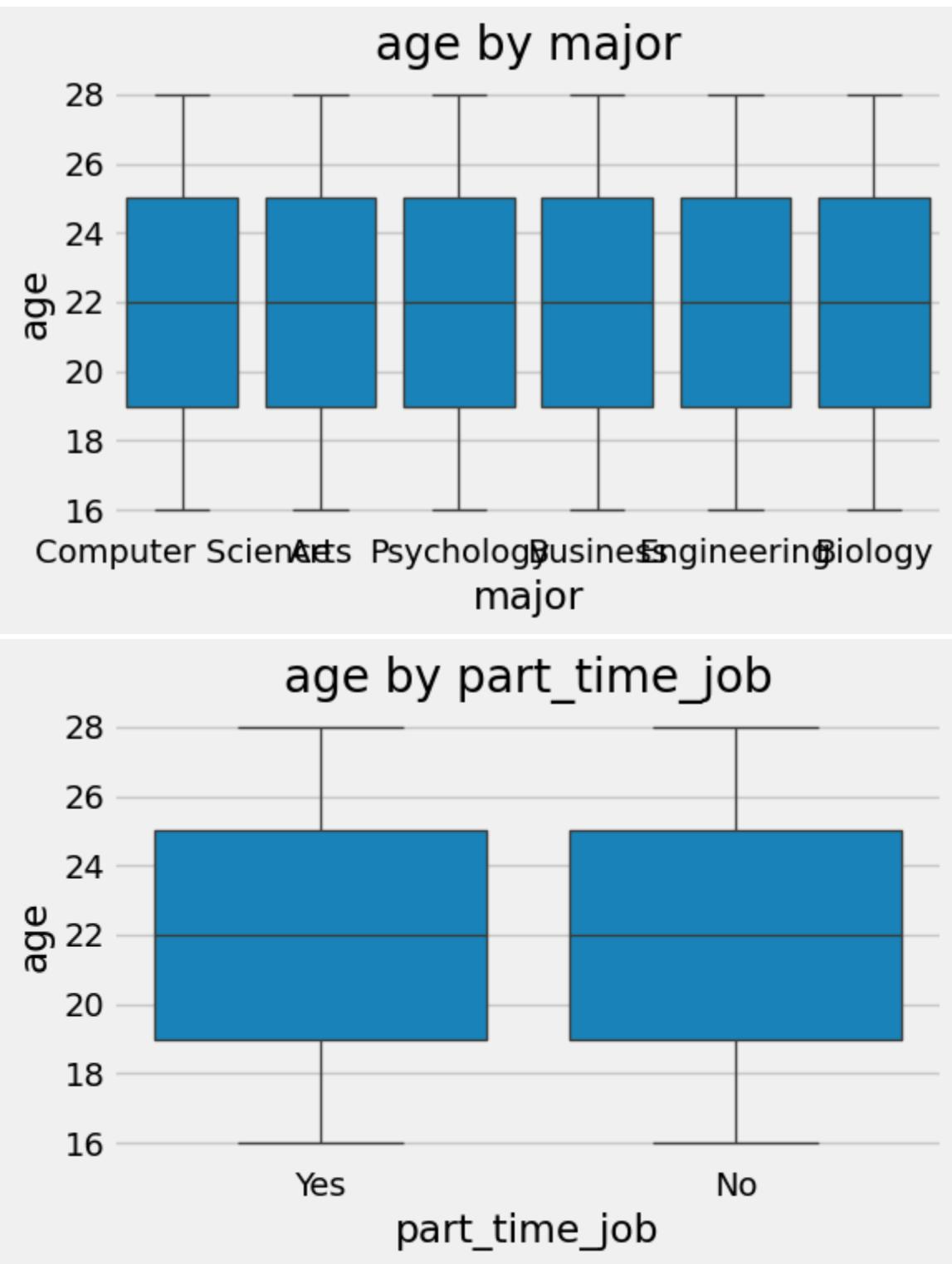
age vs exam_anxiety_score (corr=0.00)

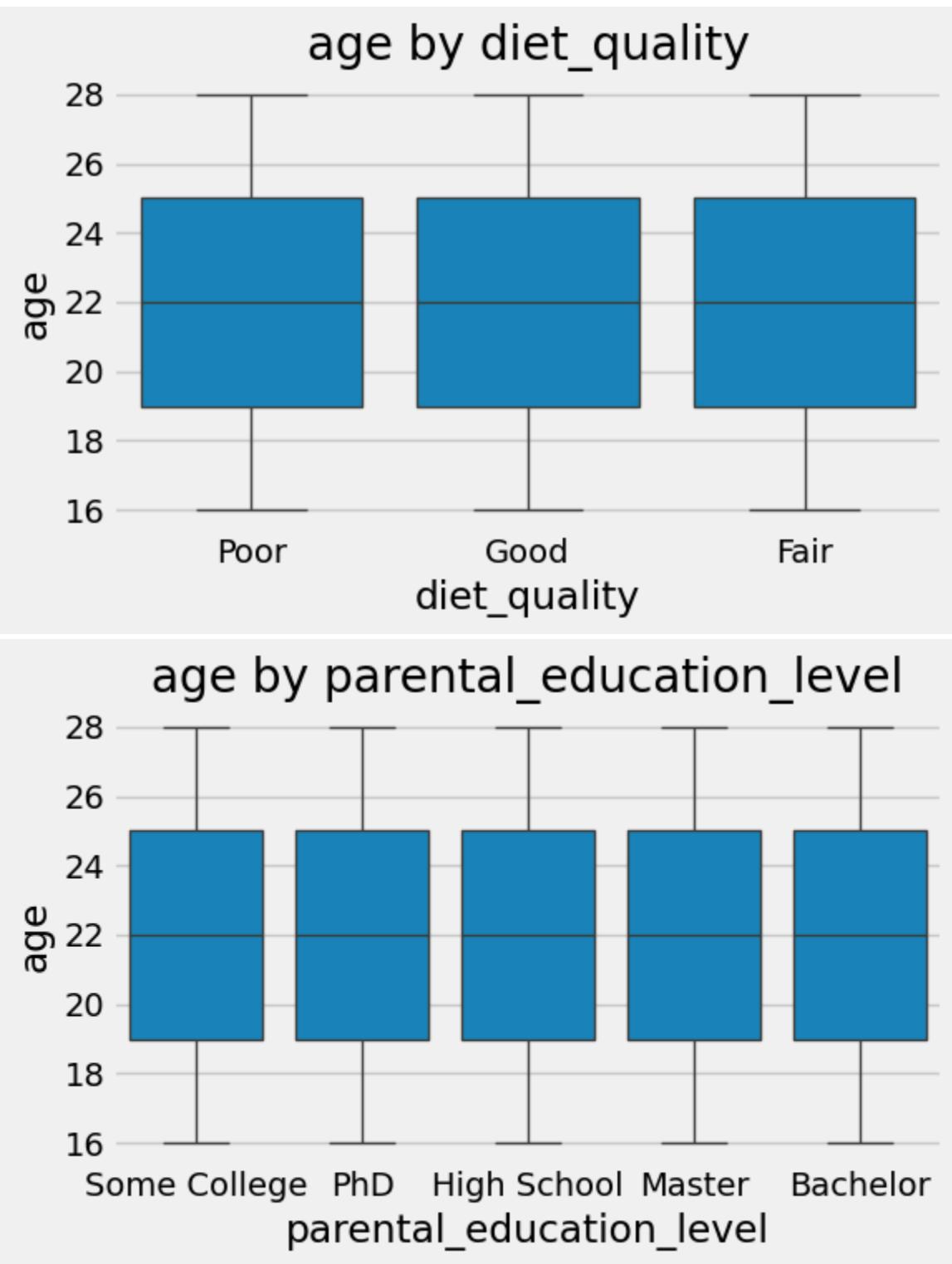


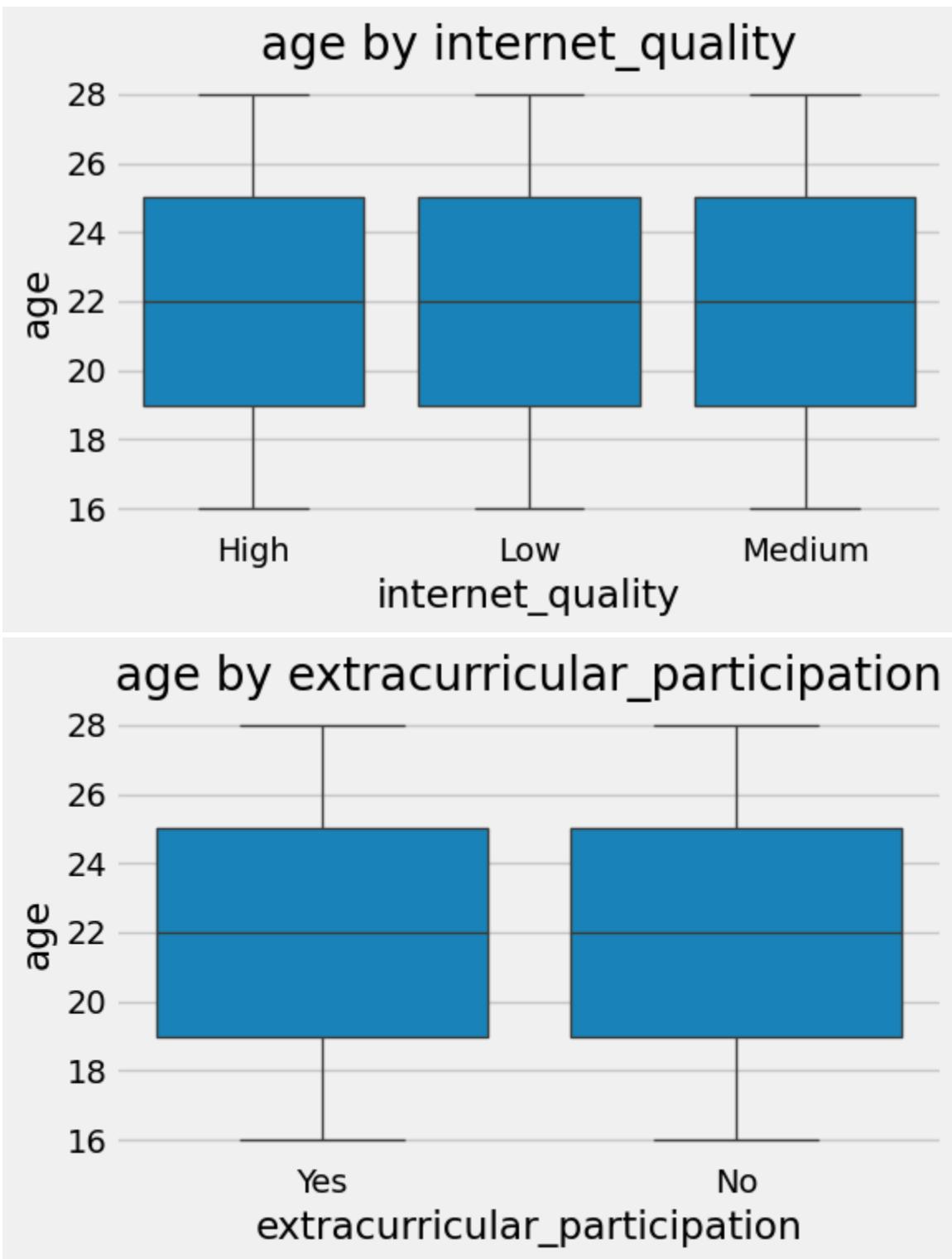
age vs time_management_score (corr=0.00)

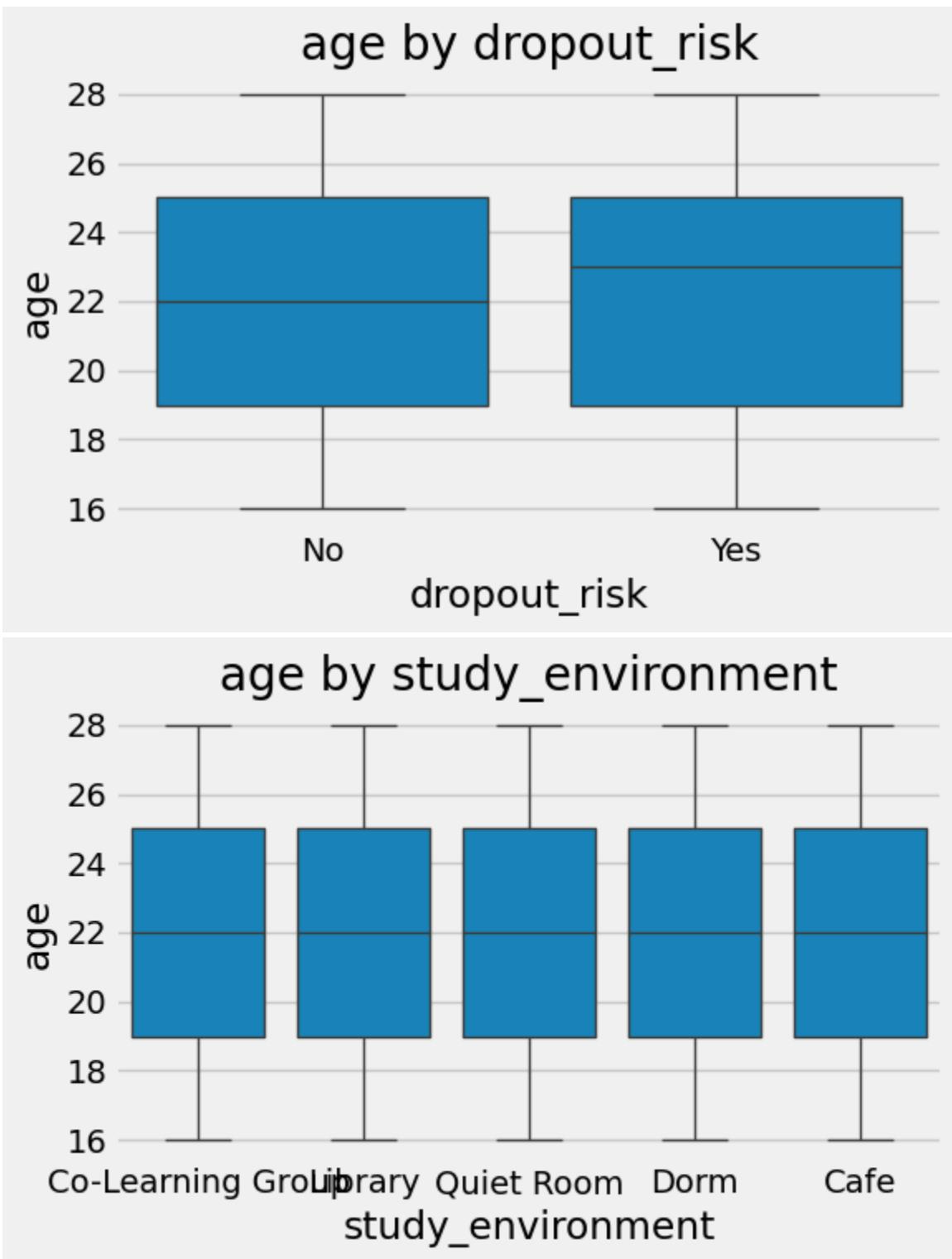


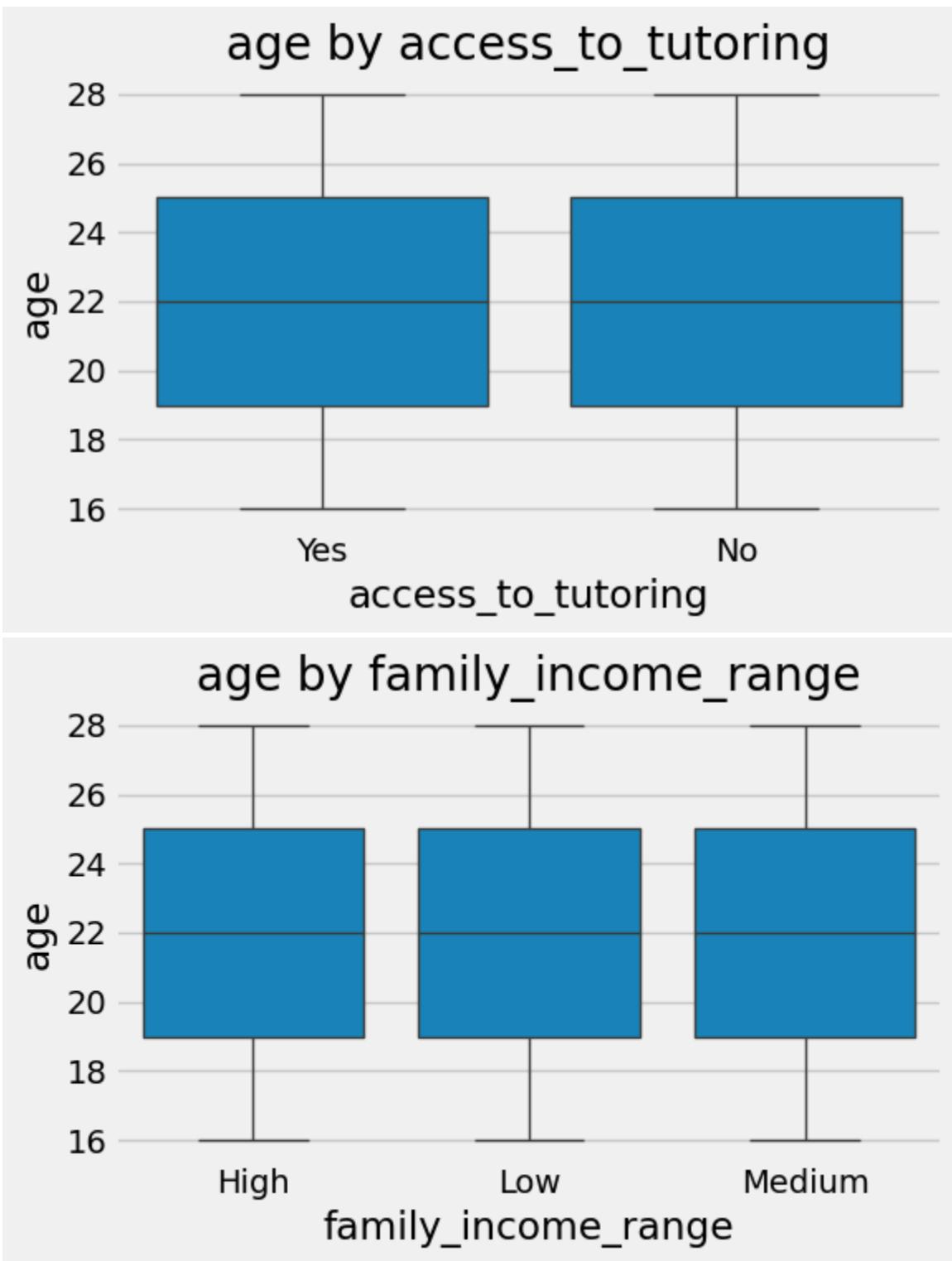


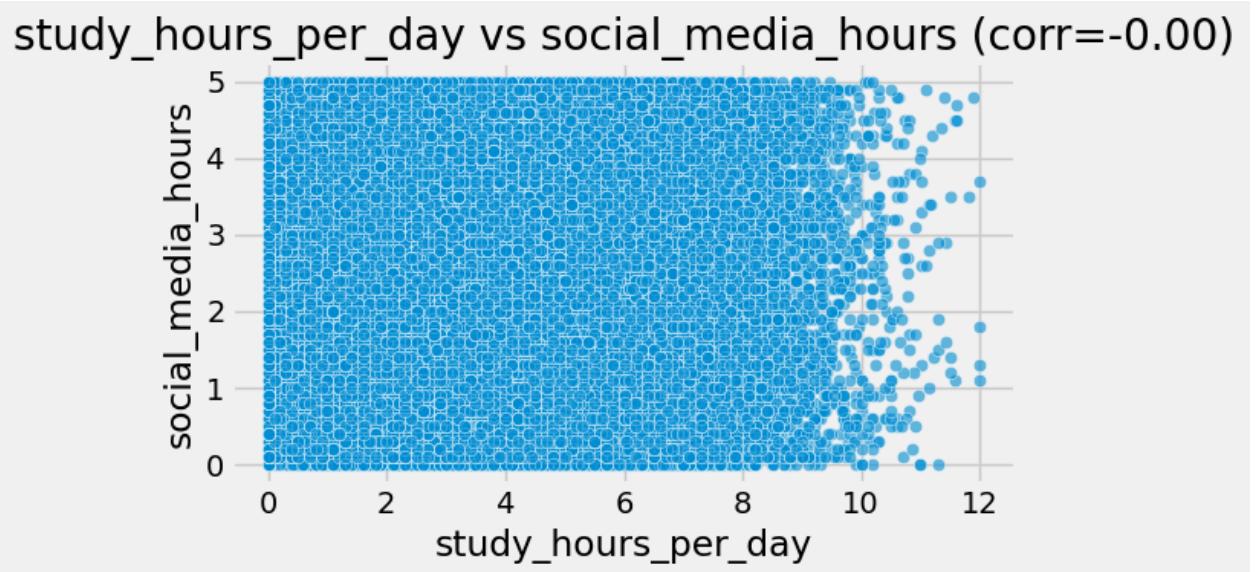
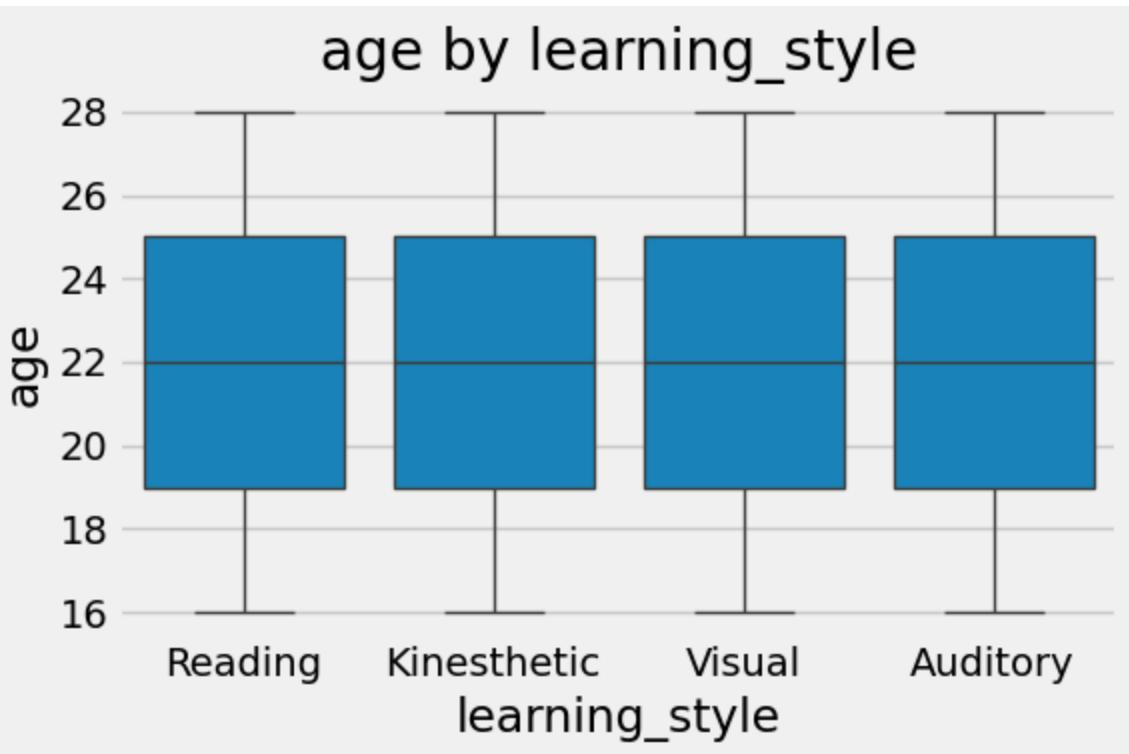




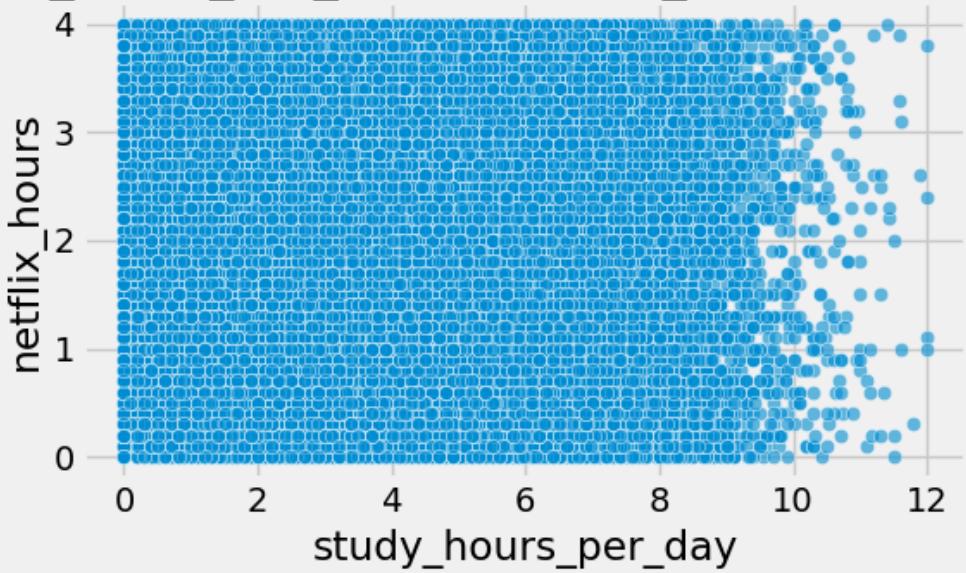




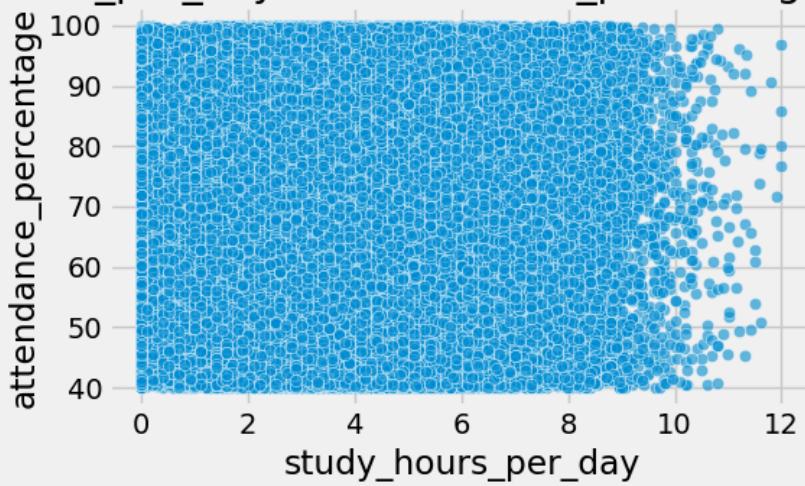




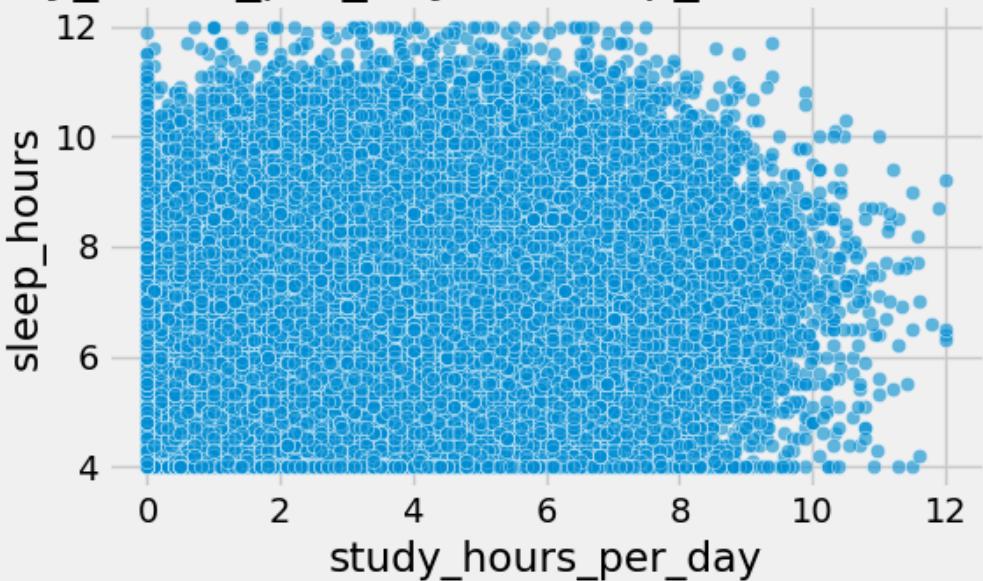
study_hours_per_day vs netflix_hours (corr=-0.00)



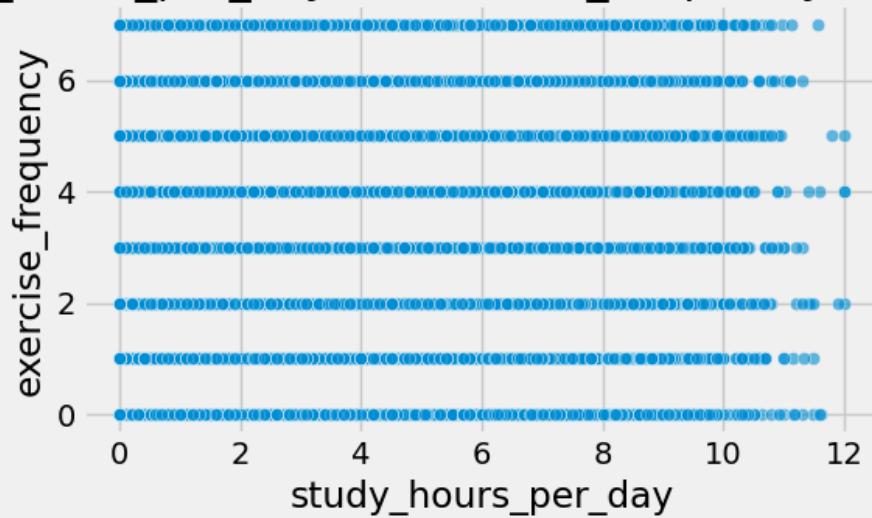
study_hours_per_day vs attendance_percentage (corr=0.00)



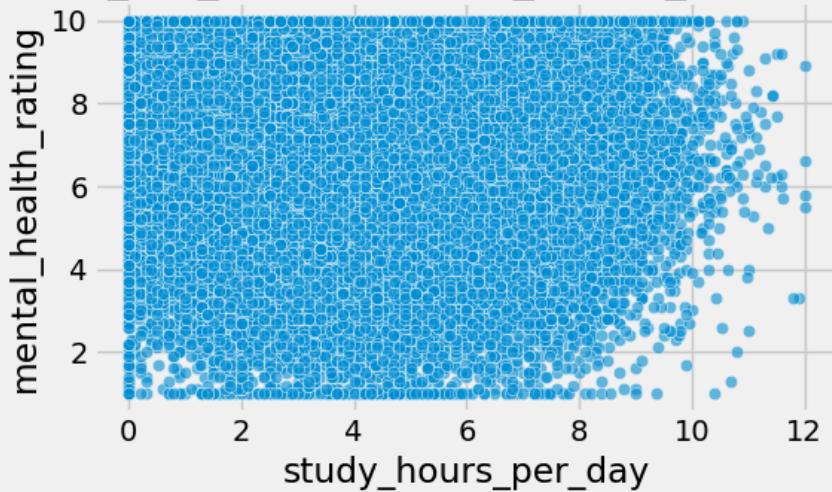
study_hours_per_day vs sleep_hours (corr=-0.00)



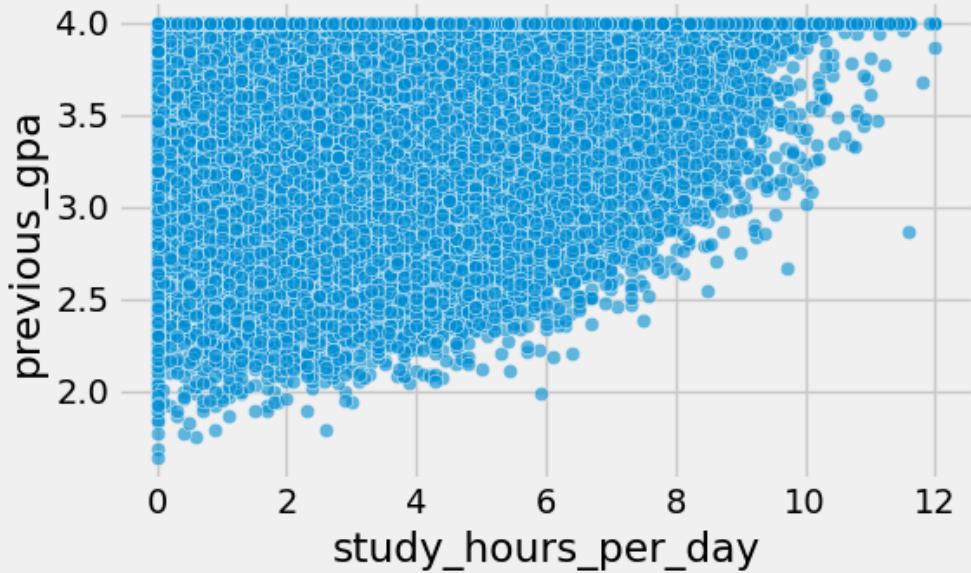
study_hours_per_day vs exercise_frequency (corr=0.00)



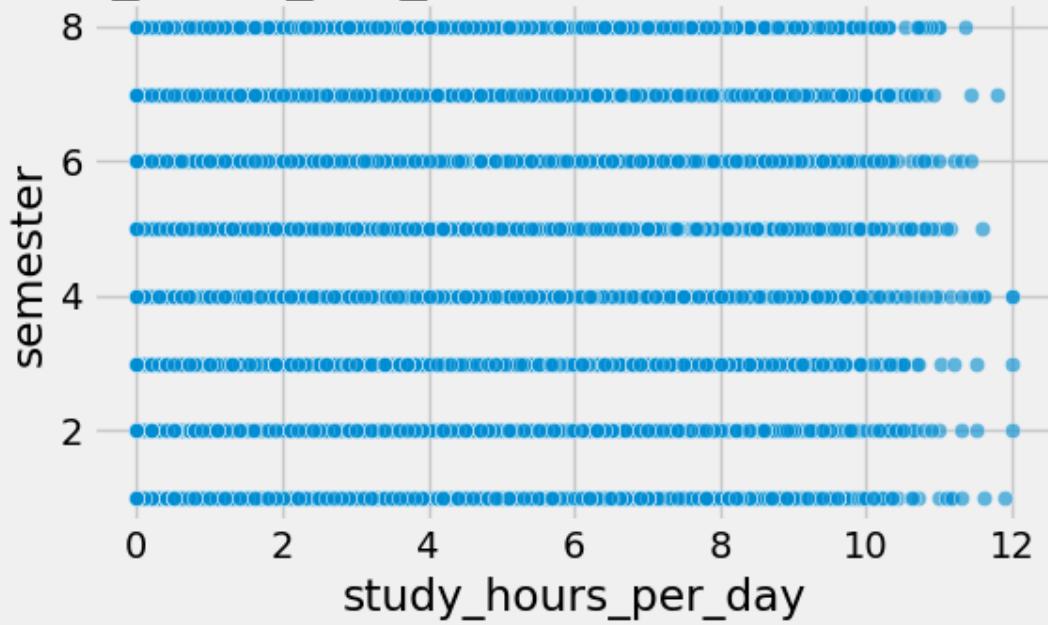
study_hours_per_day vs mental_health_rating (corr=-0.00)



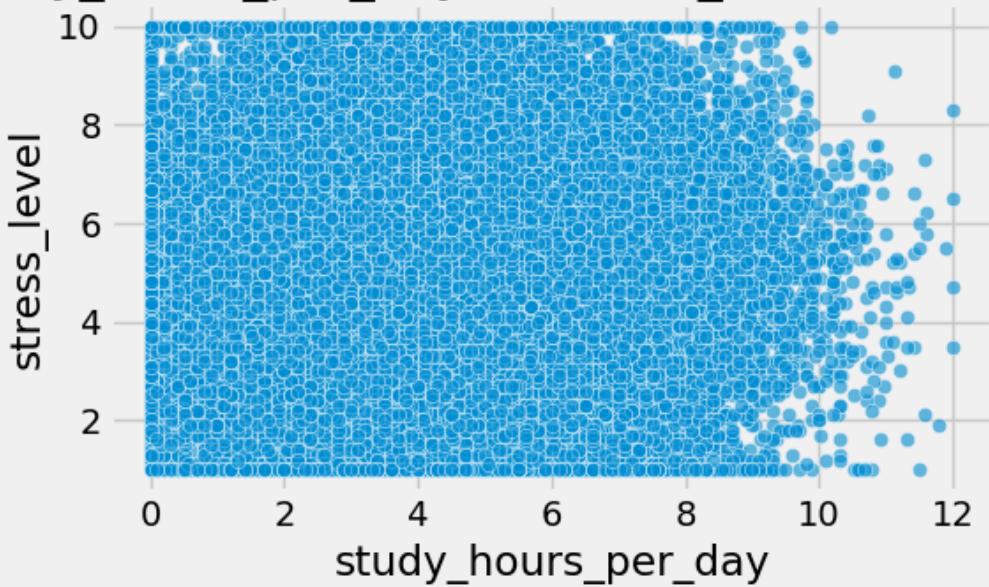
study_hours_per_day vs previous_gpa (corr=0.26)



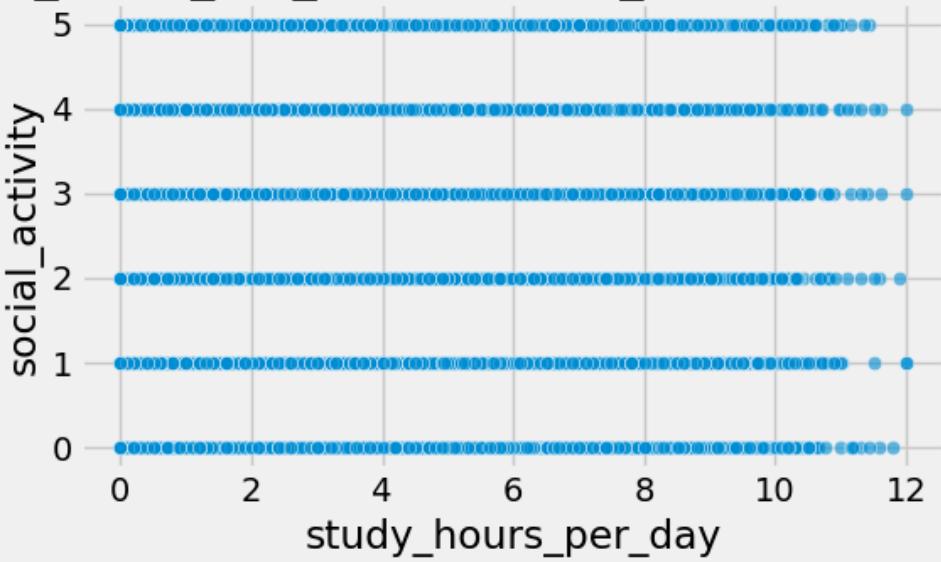
study_hours_per_day vs semester (corr=0.00)



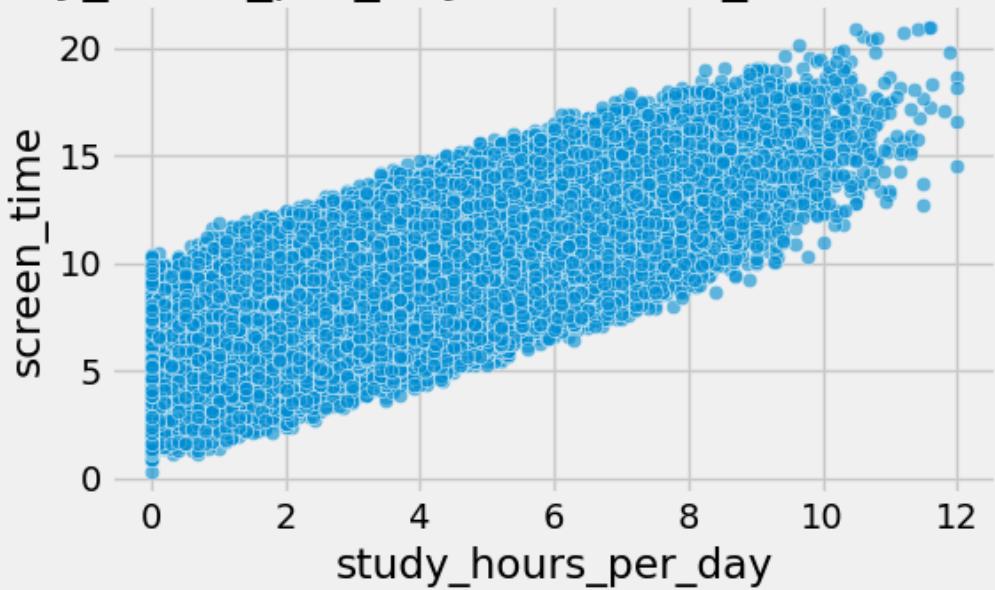
study_hours_per_day vs stress_level (corr=-0.00)



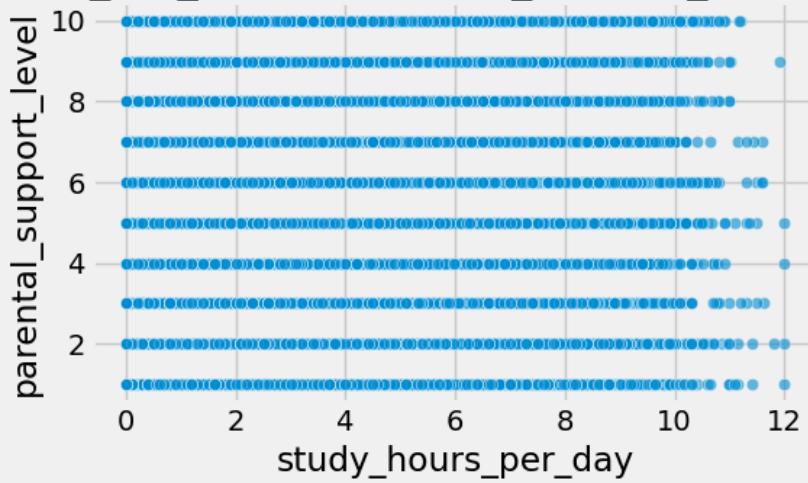
study_hours_per_day vs social_activity (corr=-0.00)



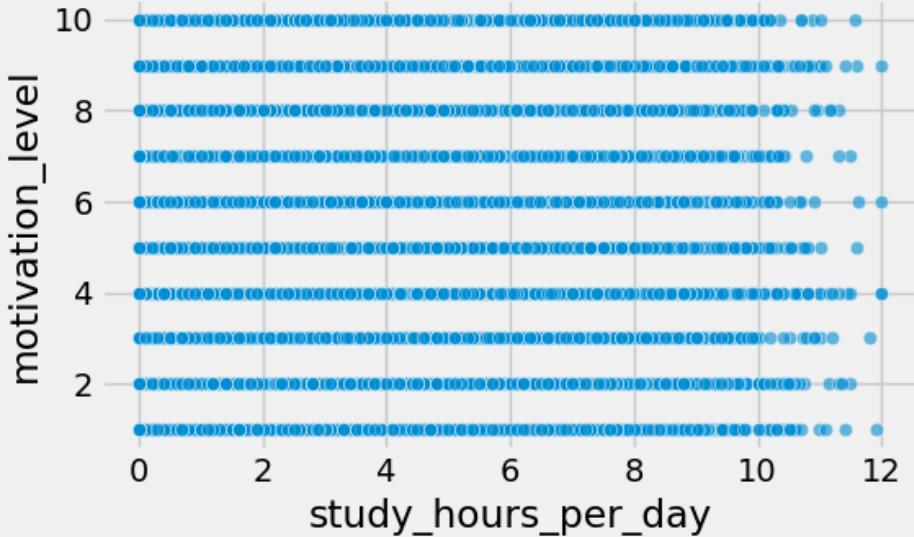
study_hours_per_day vs screen_time (corr=0.72)



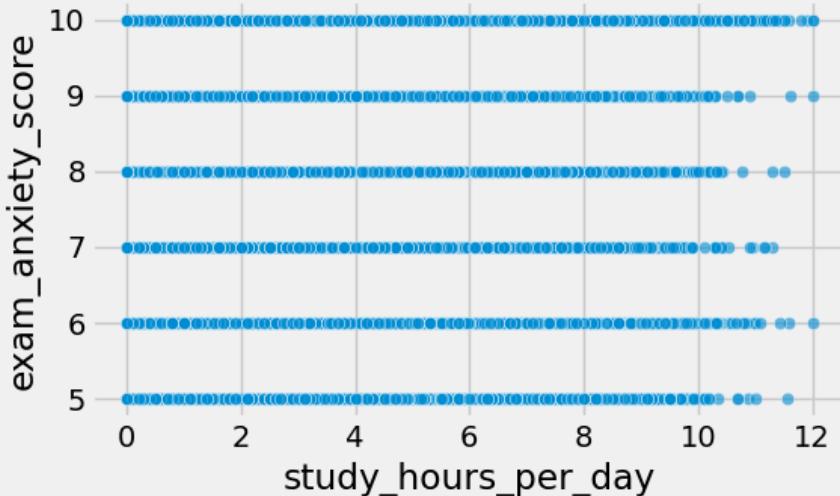
study_hours_per_day vs parental_support_level (corr=-0.00)



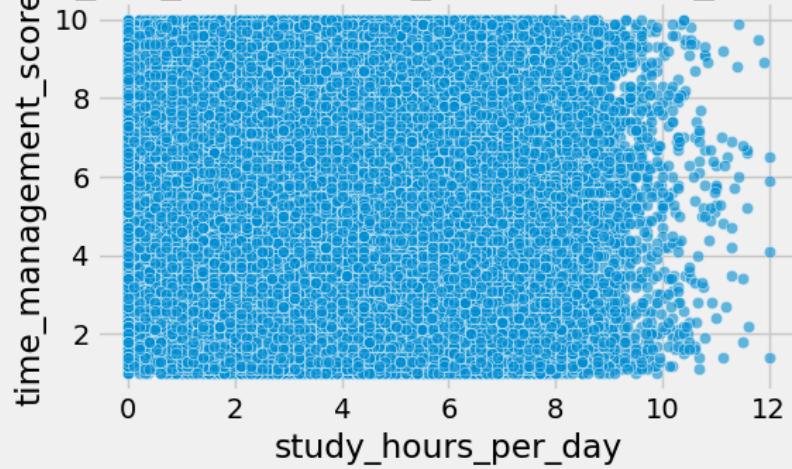
study_hours_per_day vs motivation_level (corr=0.00)



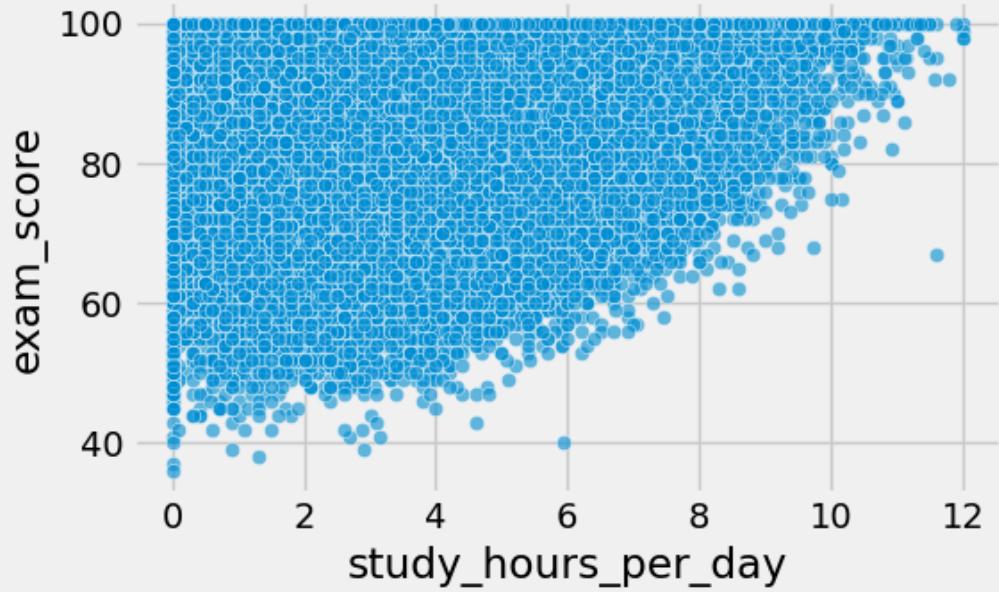
study_hours_per_day vs exam_anxiety_score (corr=-0.00)



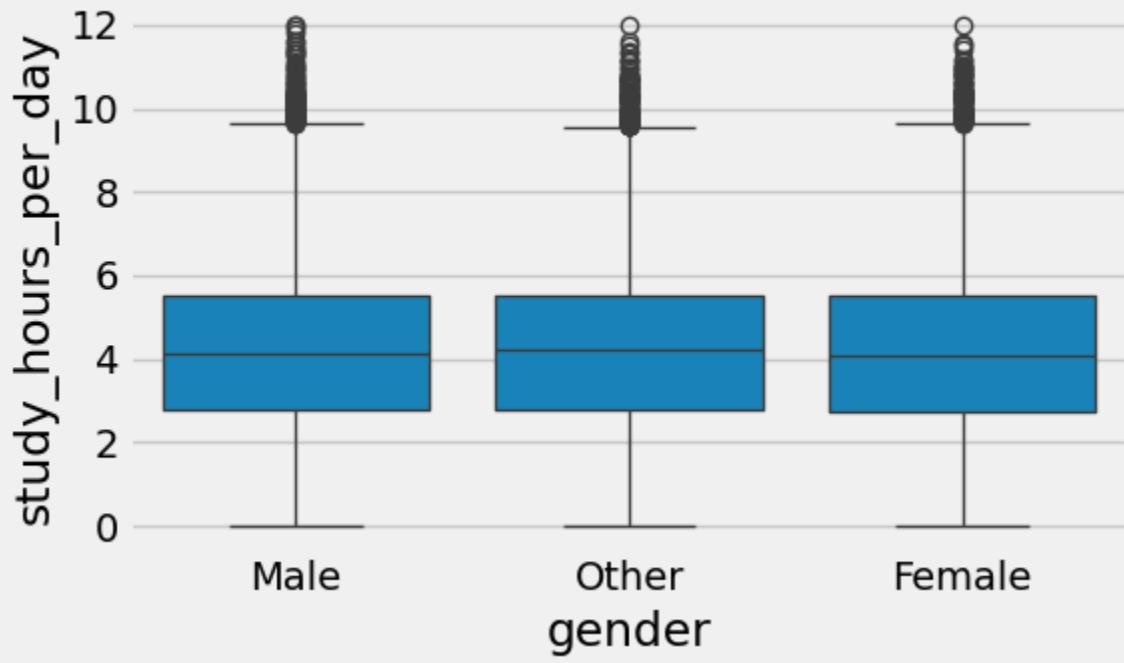
study_hours_per_day vs time_management_score (corr=0.00)

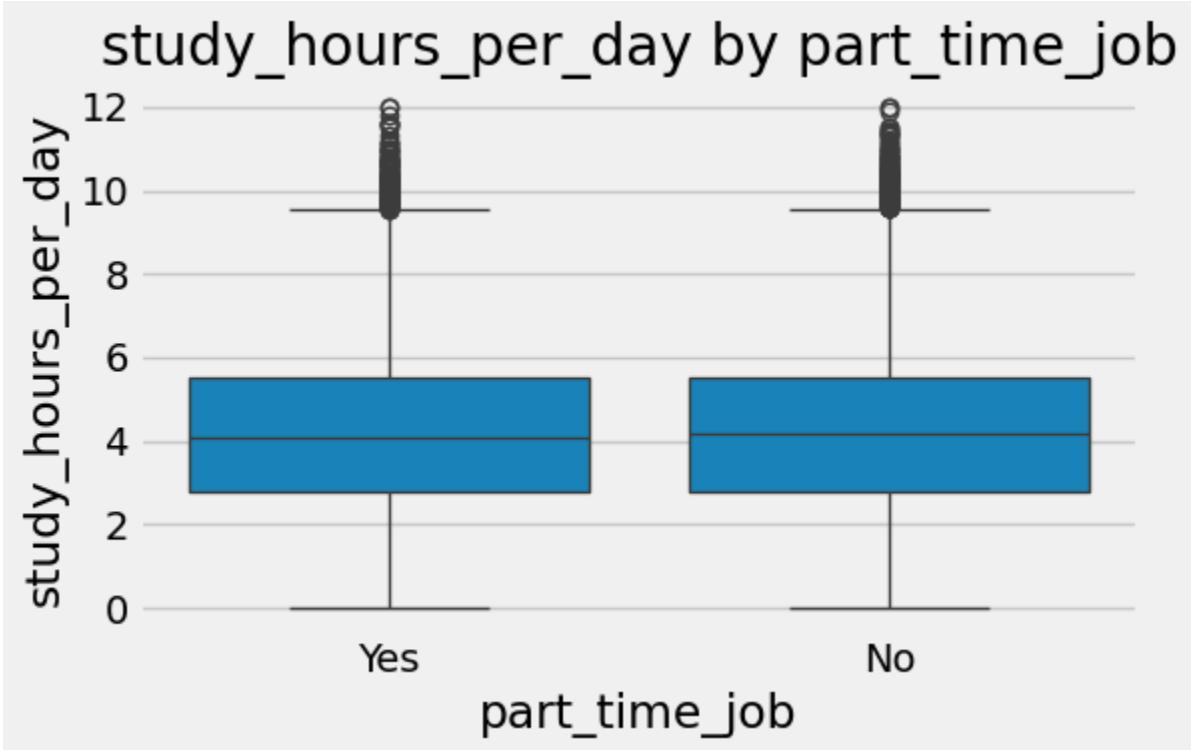
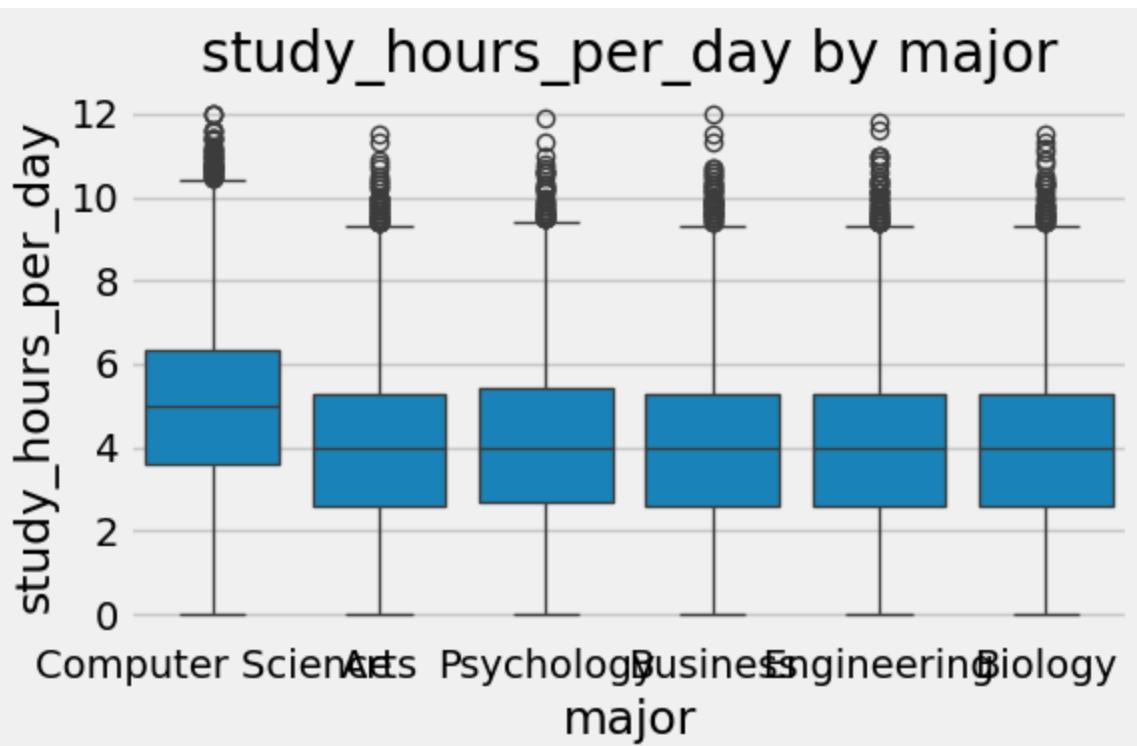


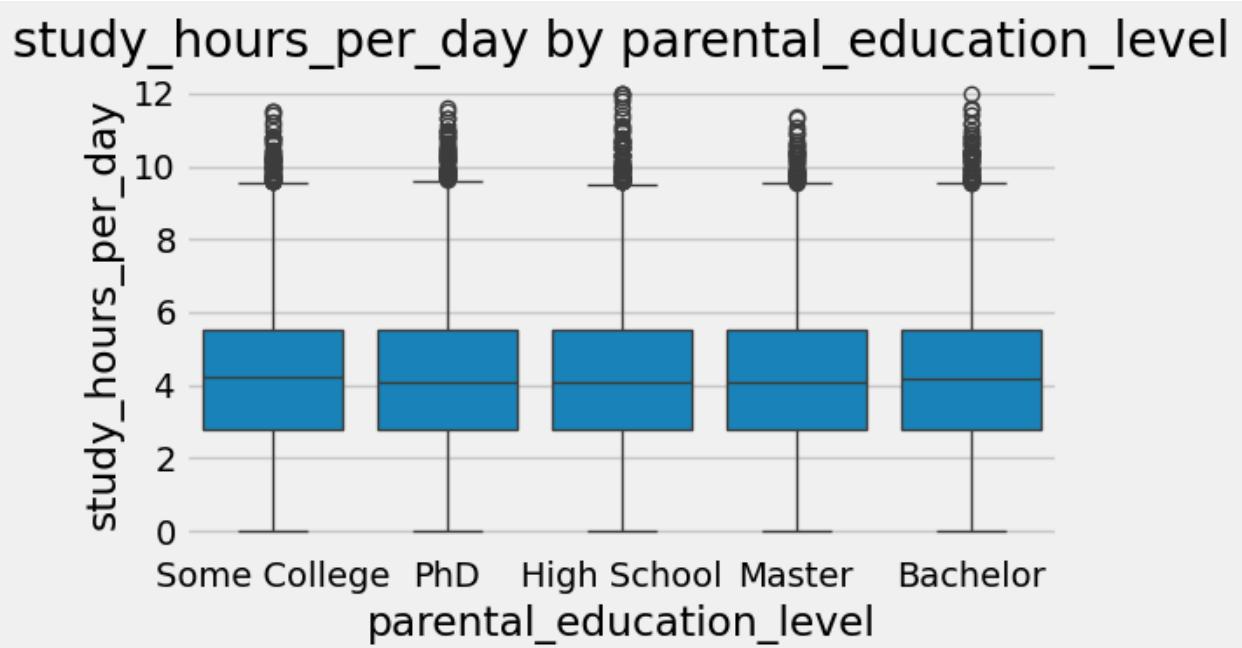
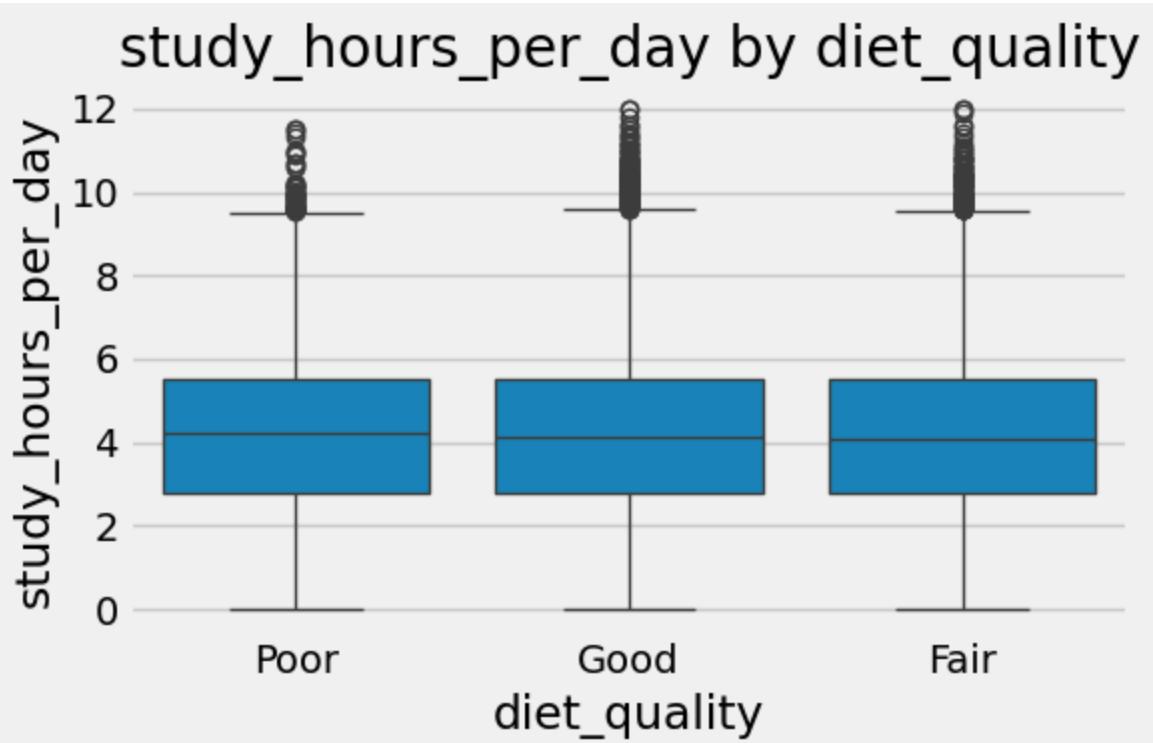
study_hours_per_day vs exam_score (corr=0.24)

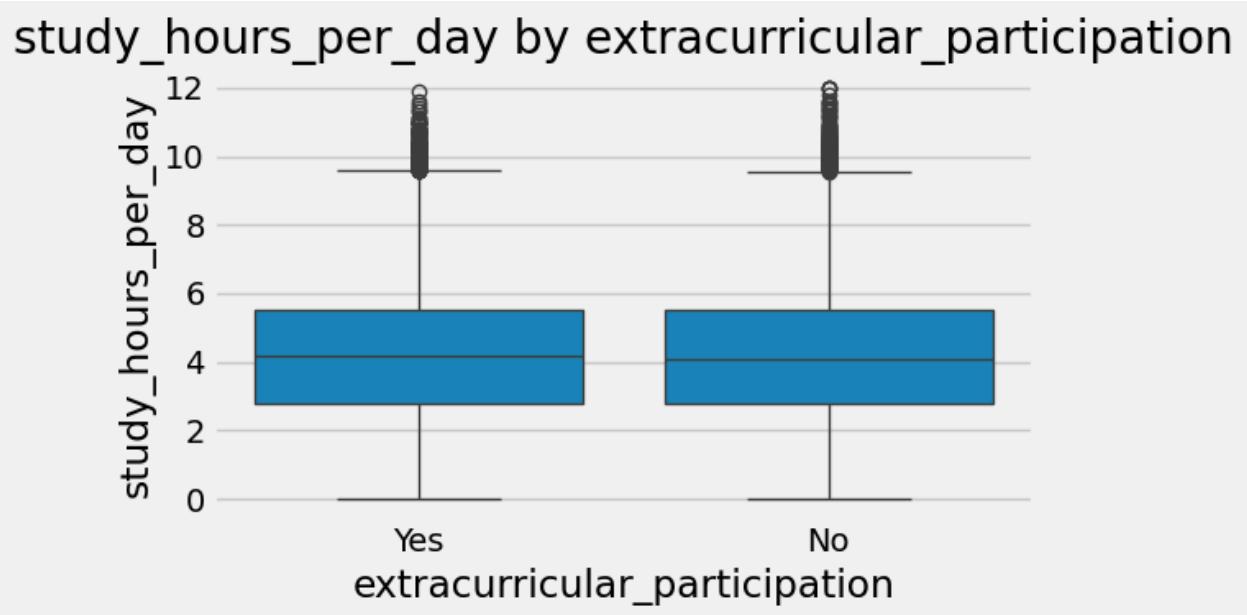
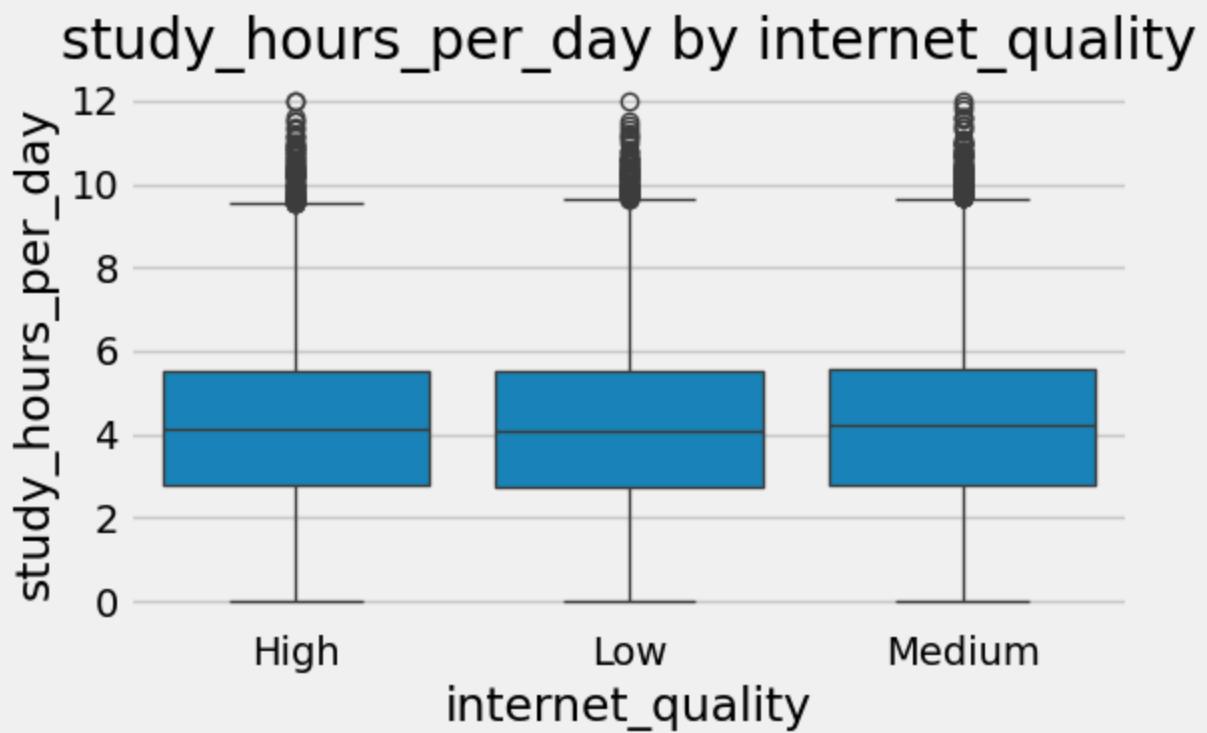


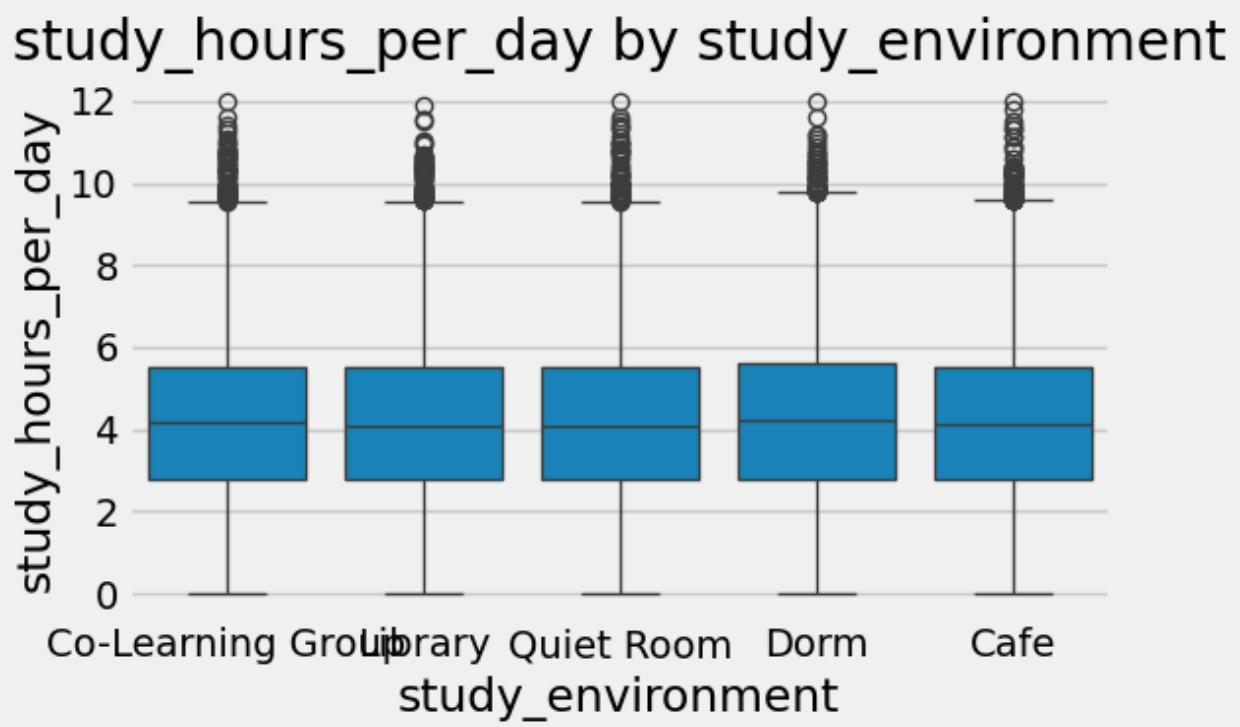
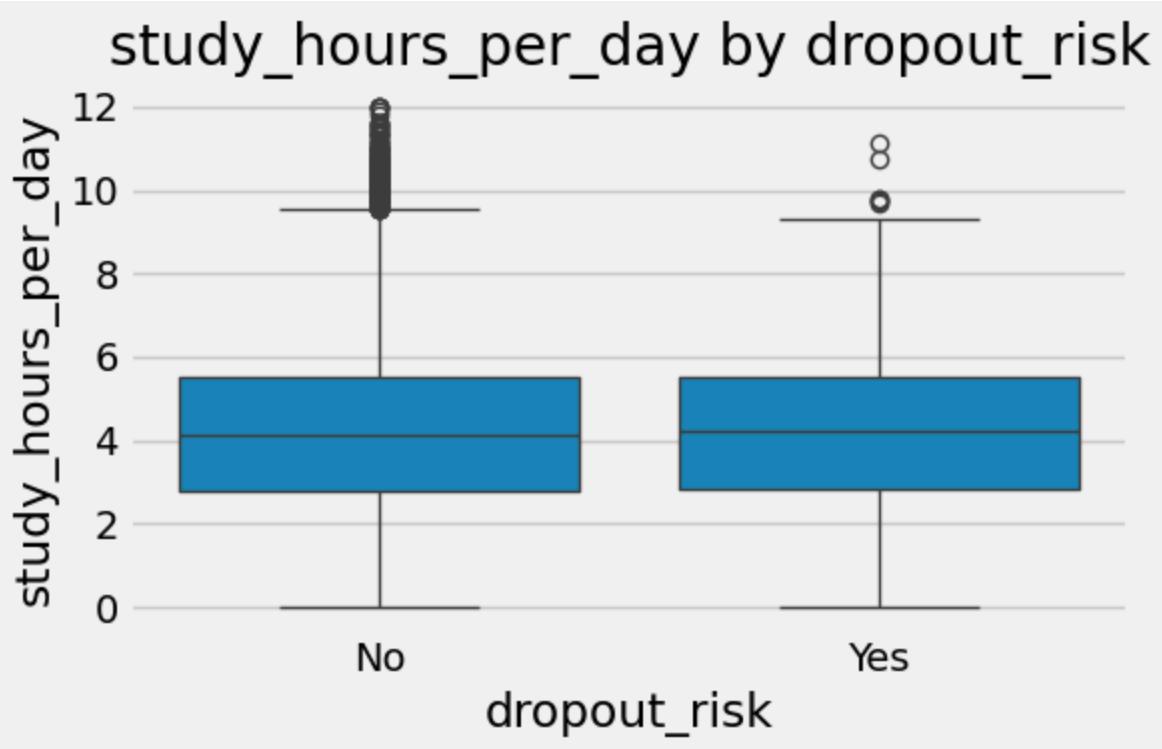
study_hours_per_day by gender

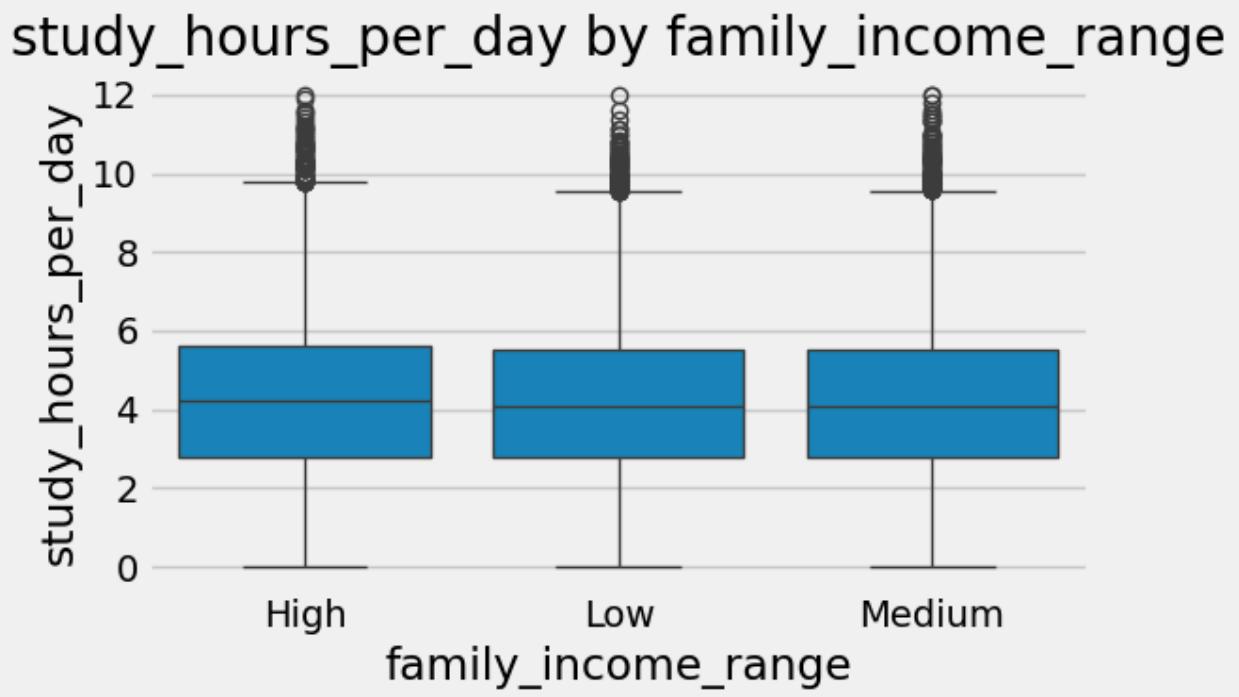
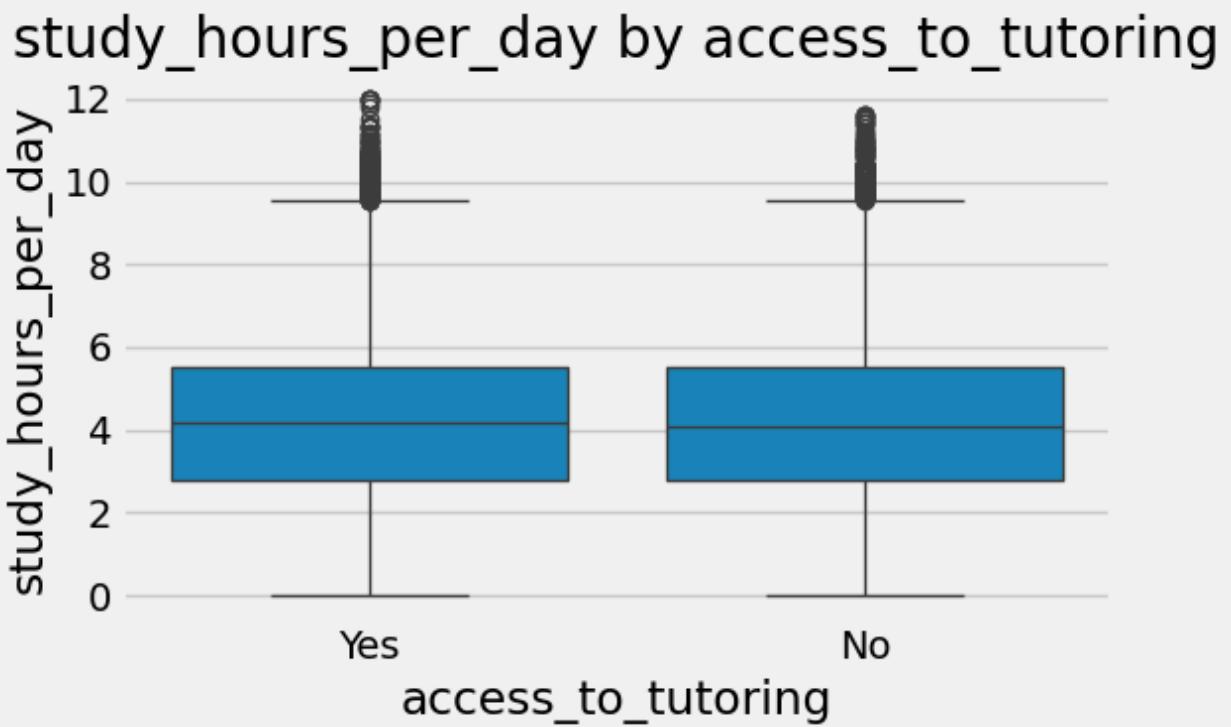


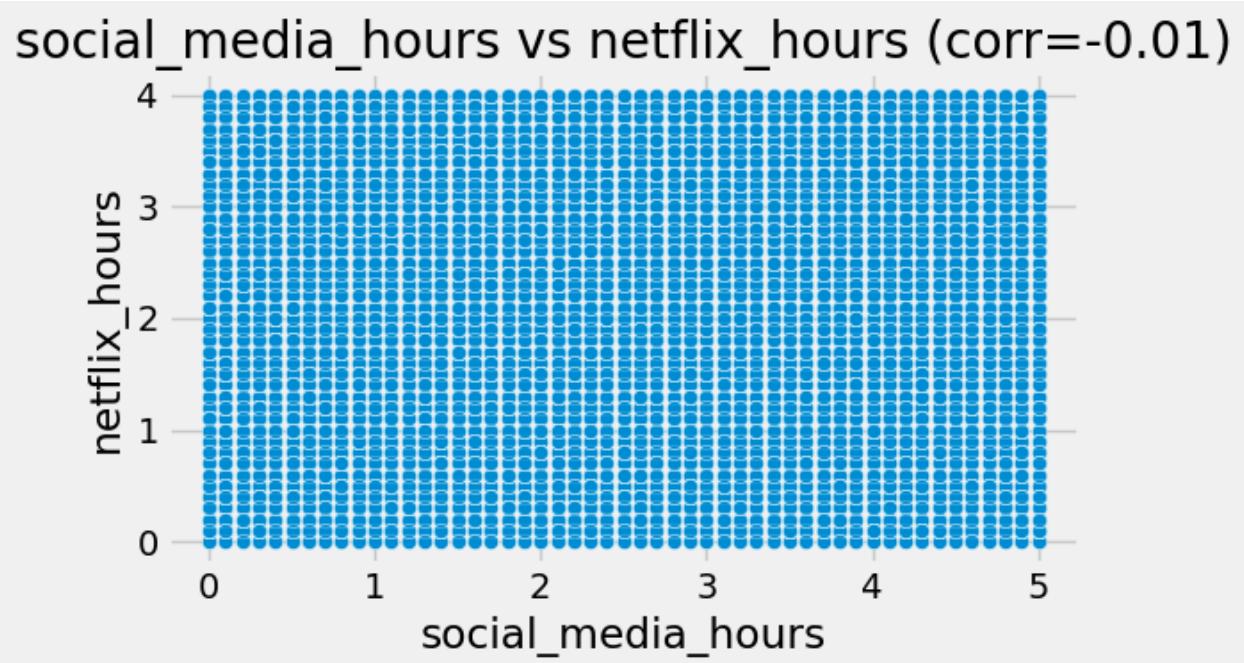
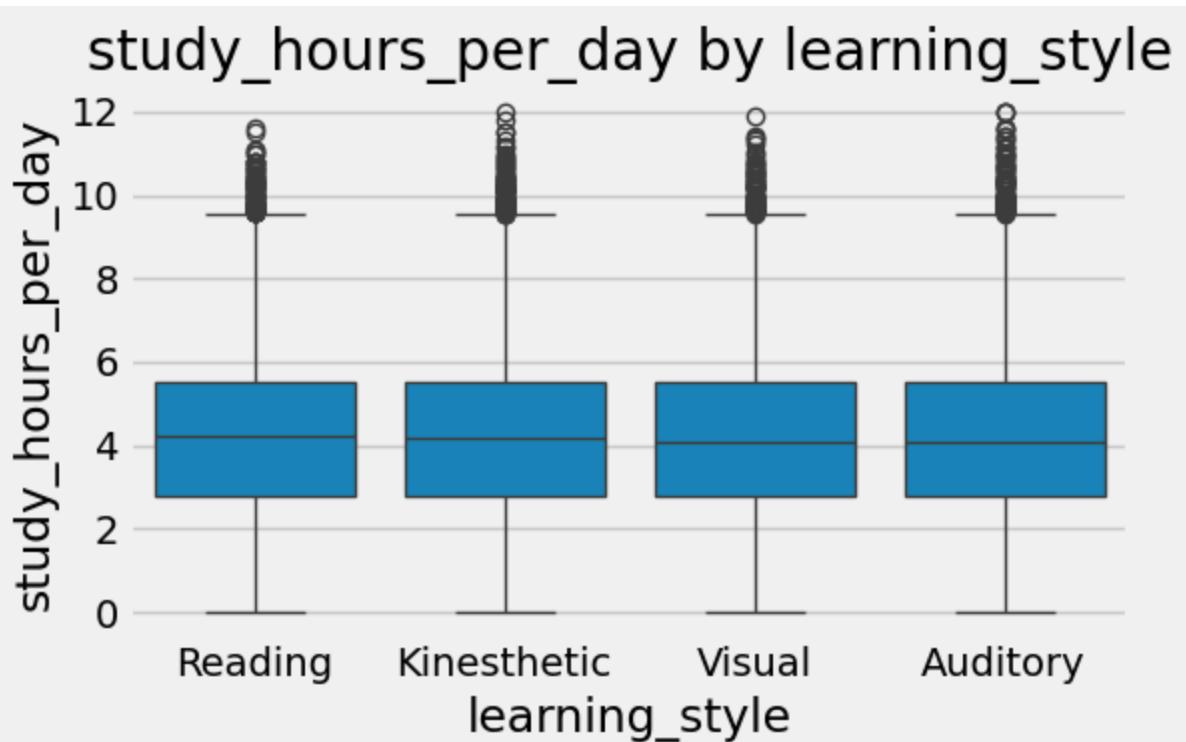




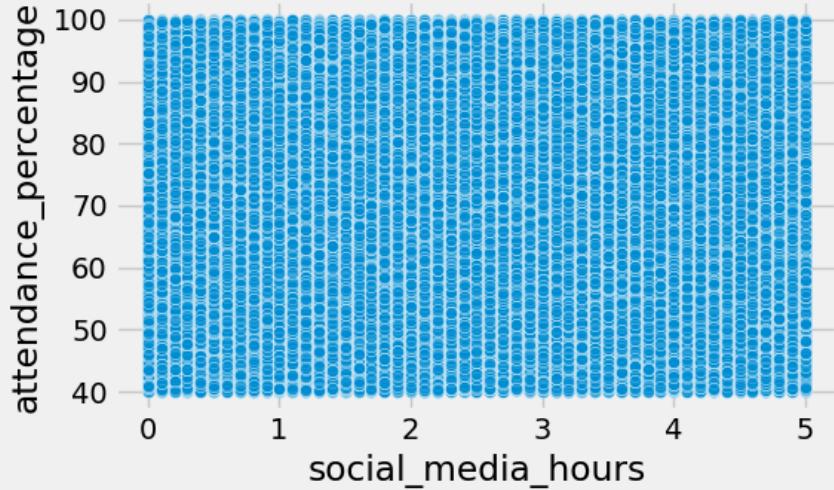




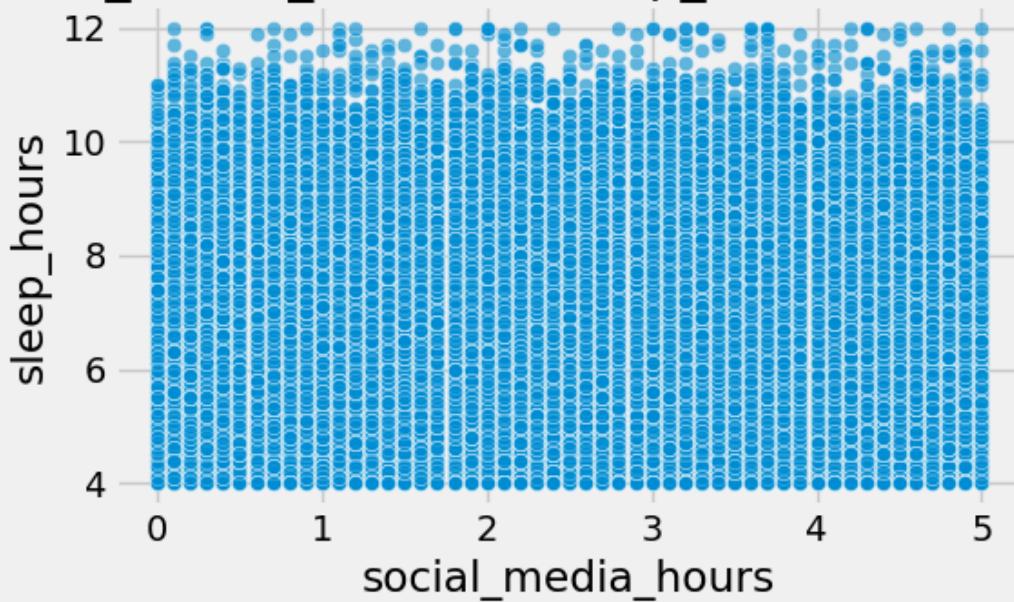




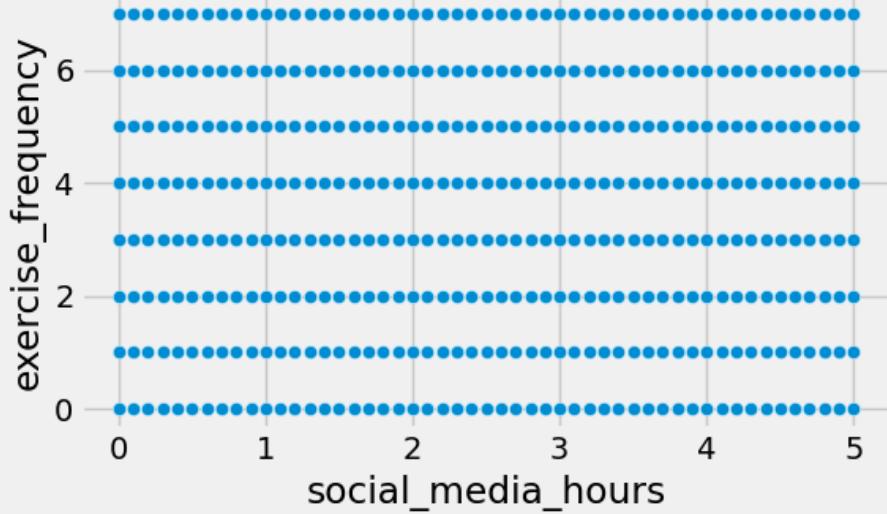
`social_media_hours` vs `attendance_percentage` (corr=0.00)



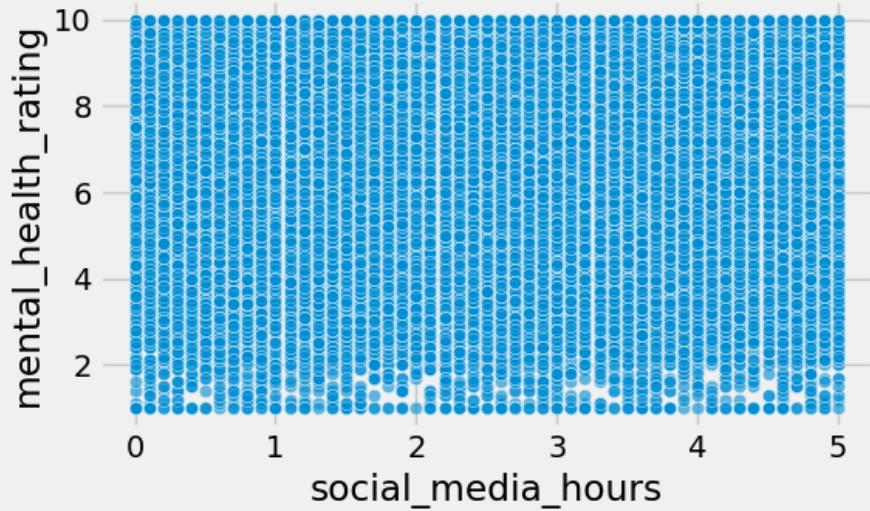
`social_media_hours` vs `sleep_hours` (corr=-0.00)



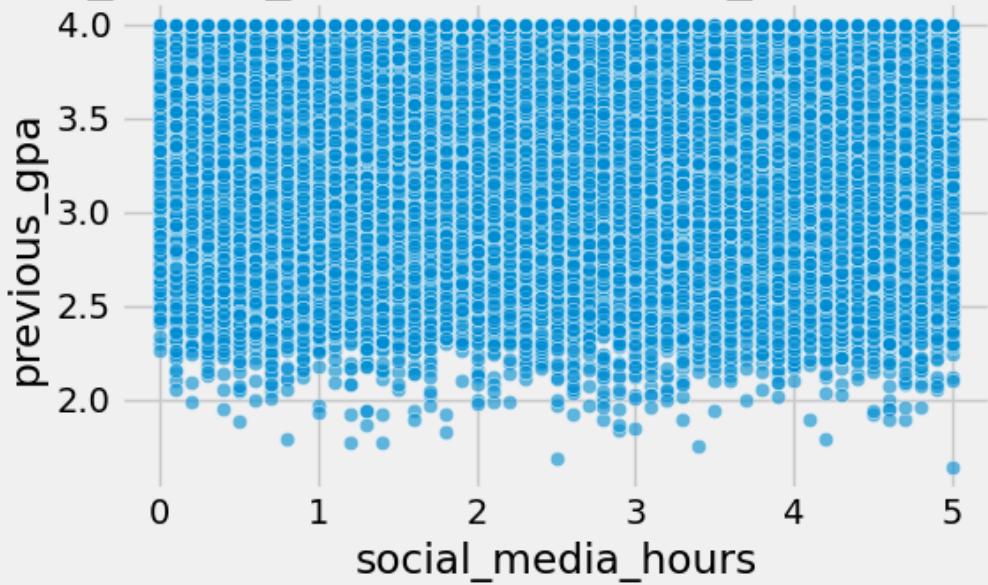
social_media_hours vs exercise_frequency (corr=-0.01)



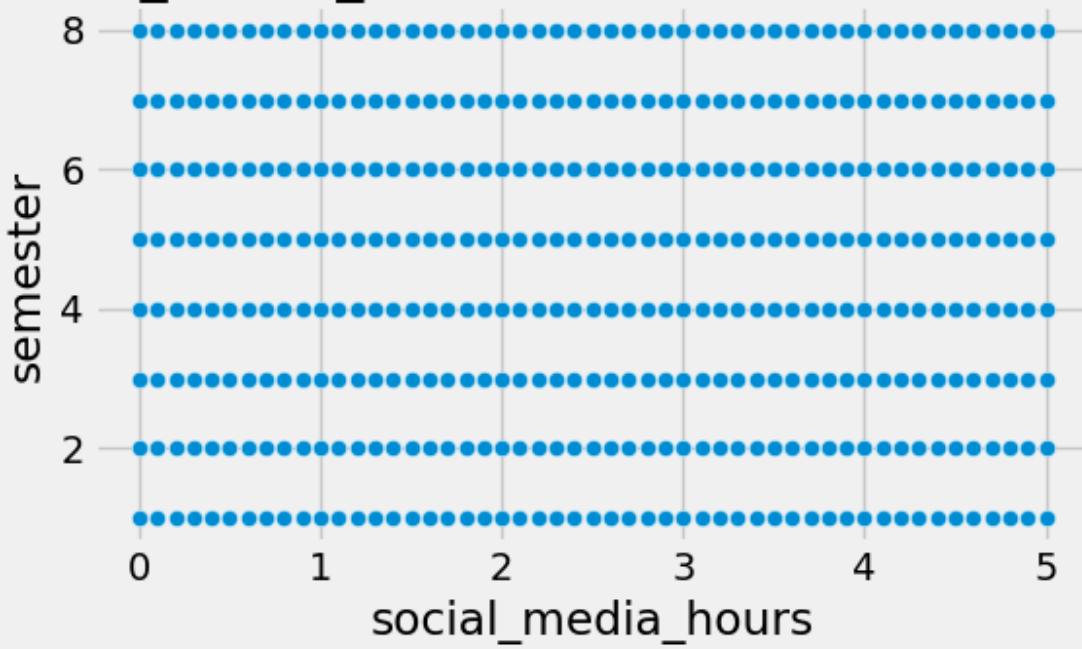
social_media_hours vs mental_health_rating (corr=0.00)



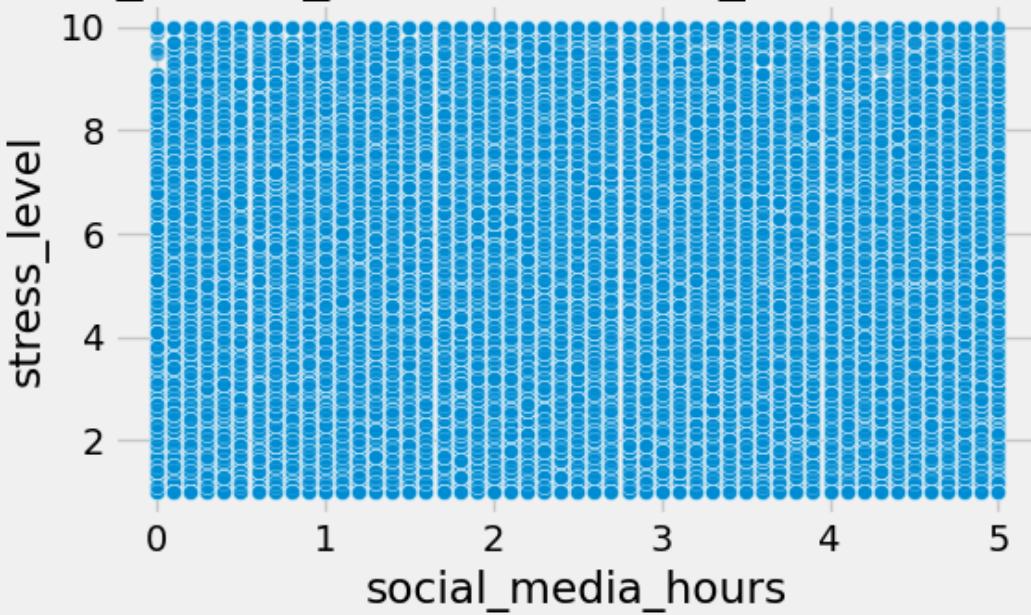
social_media_hours vs previous_gpa (corr=-0.01)



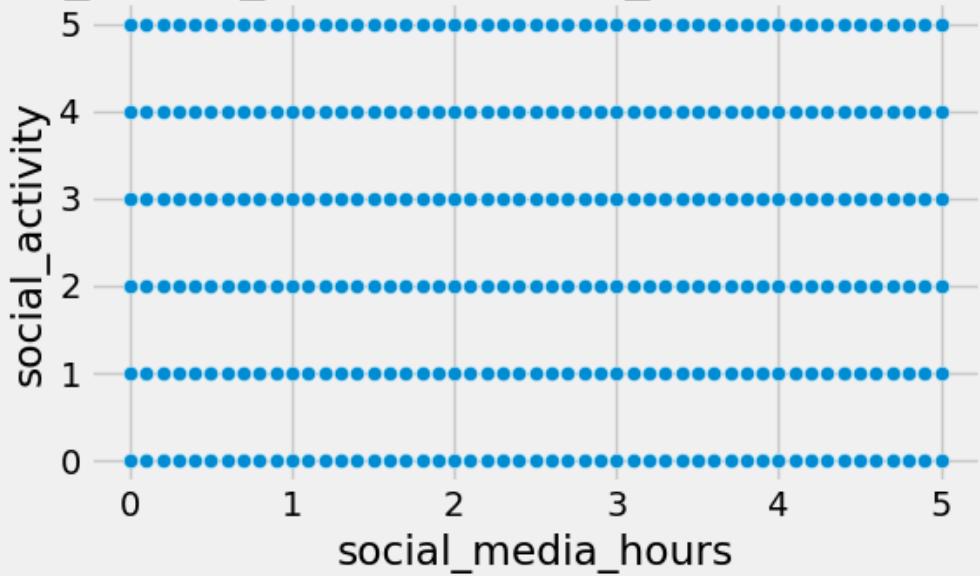
social_media_hours vs semester (corr=0.00)



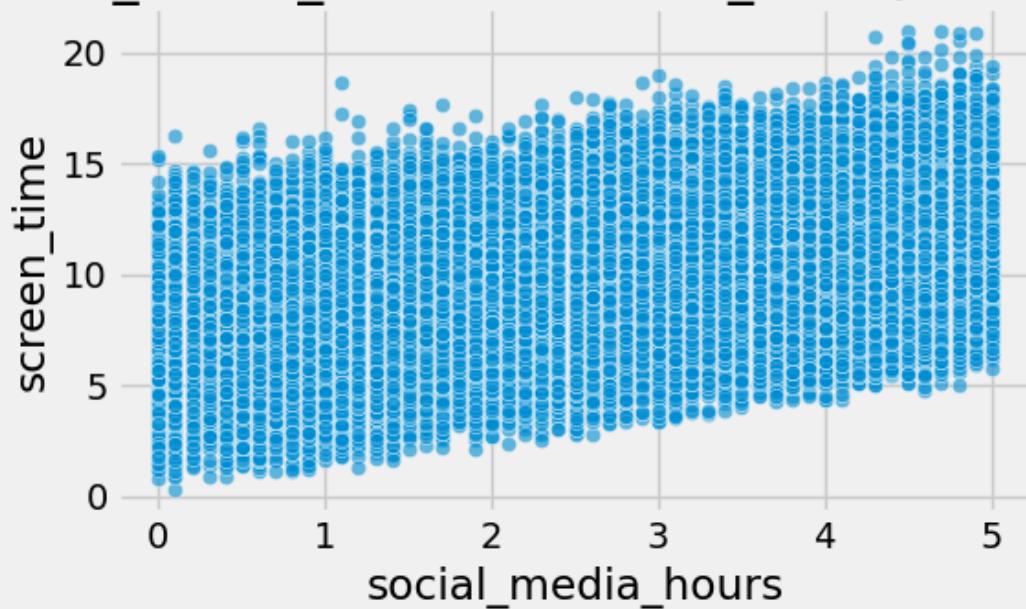
social_media_hours vs stress_level (corr=0.00)



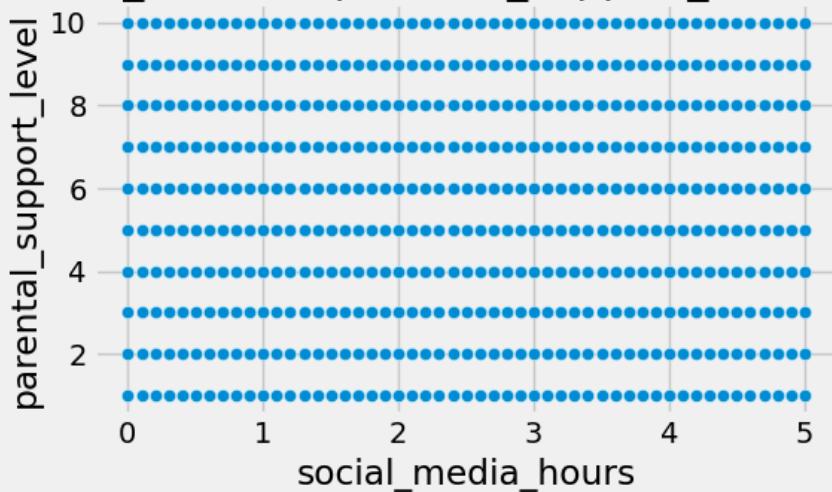
social_media_hours vs social_activity (corr=-0.00)



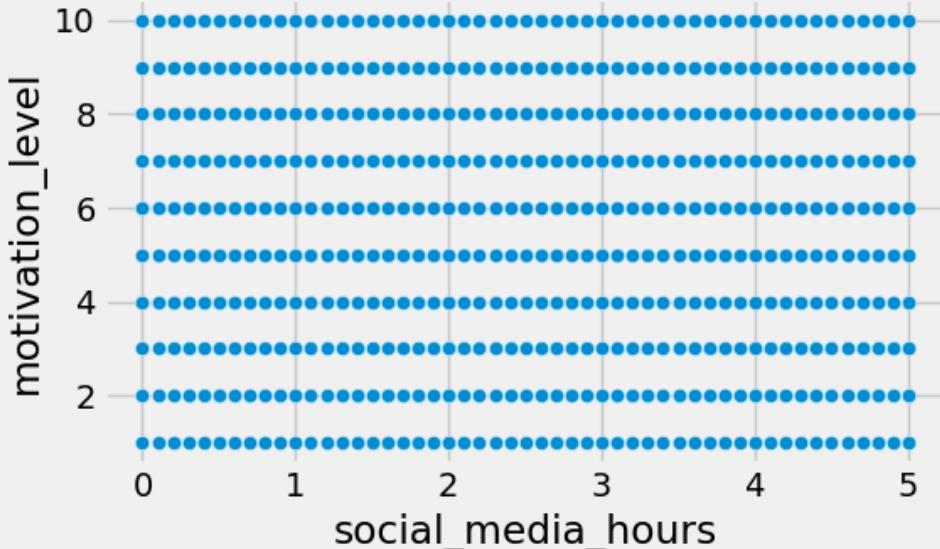
social_media_hours vs screen_time (corr=0.52)



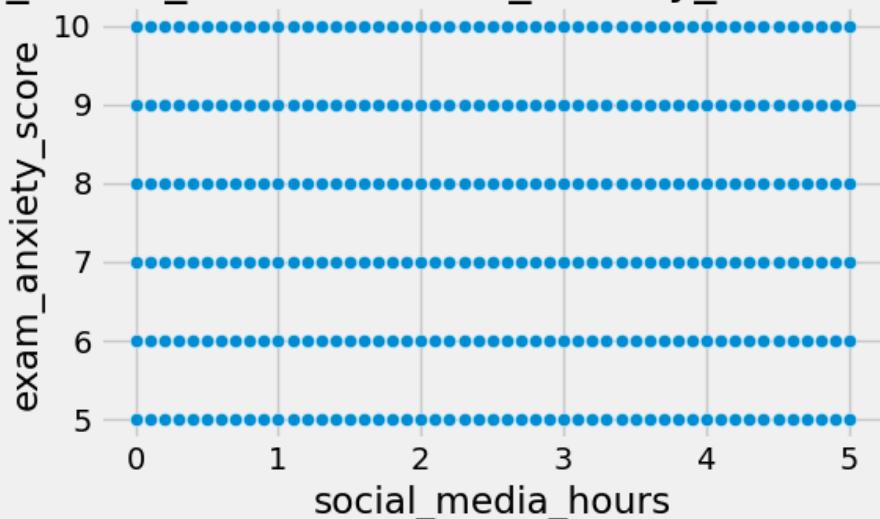
social_media_hours vs parental_support_level (corr=-0.00)



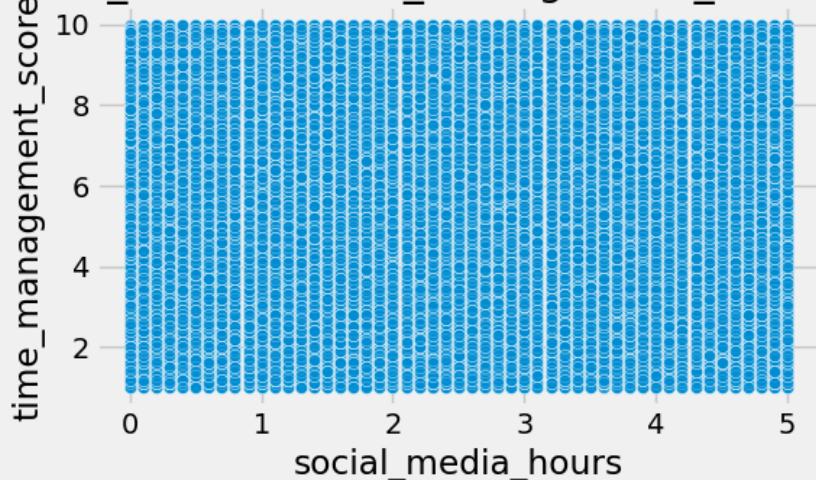
social_media_hours vs motivation_level (corr=0.00)



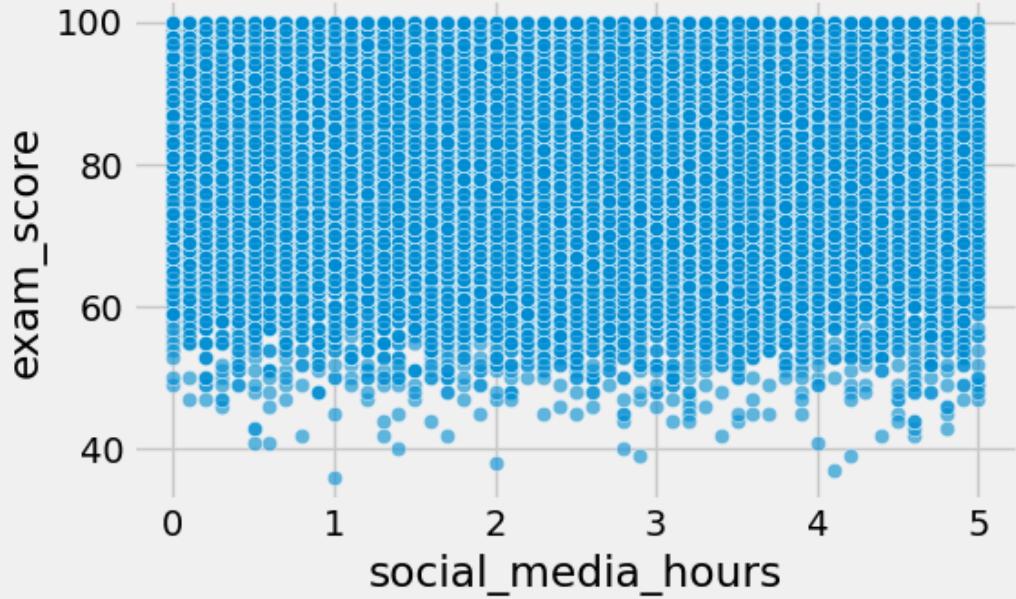
social_media_hours vs exam_anxiety_score (corr=0.00)



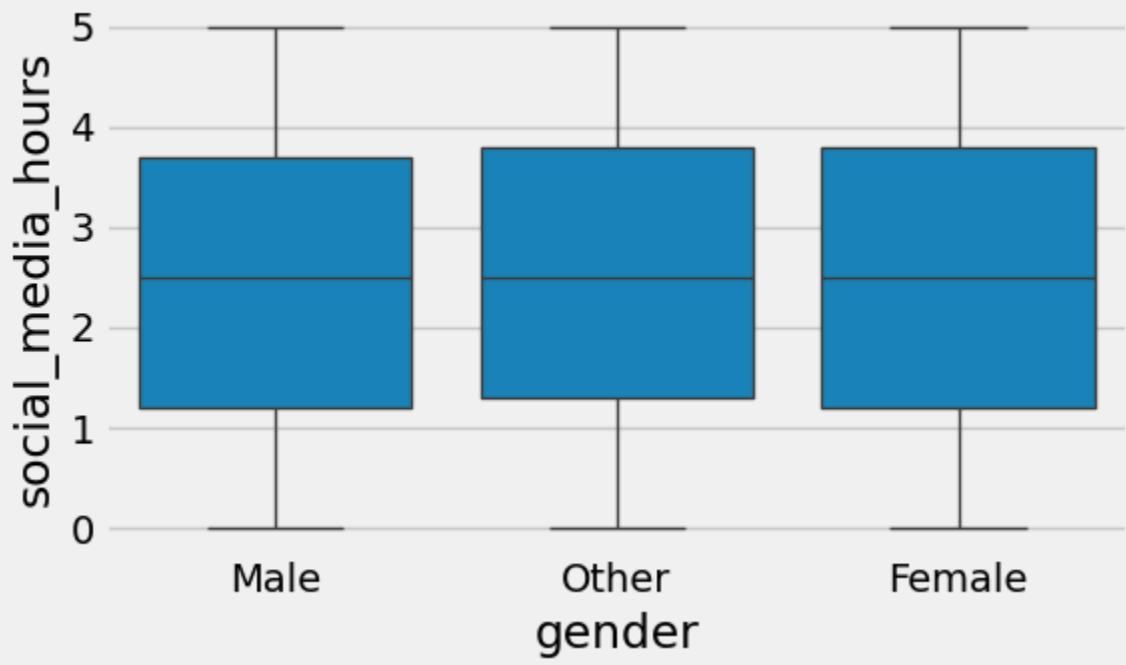
social_media_hours vs time_management_score (corr=0.00)

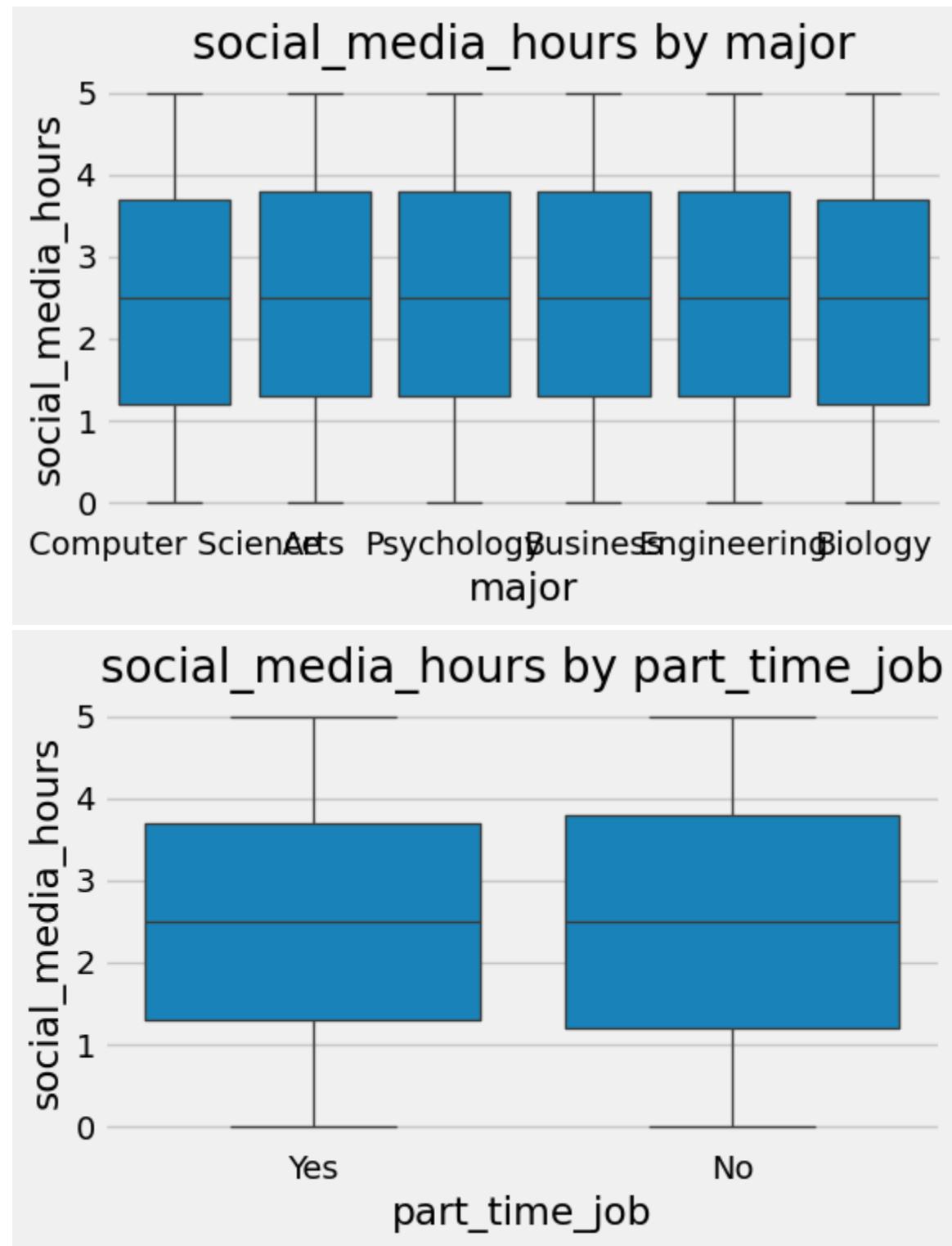


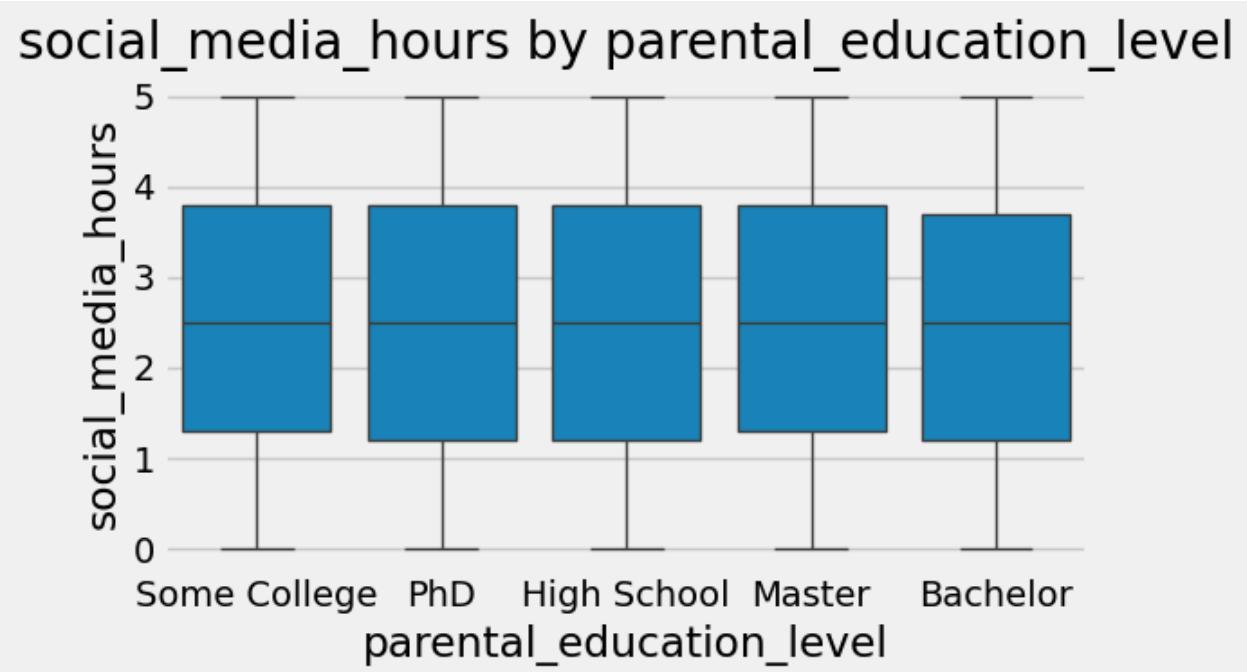
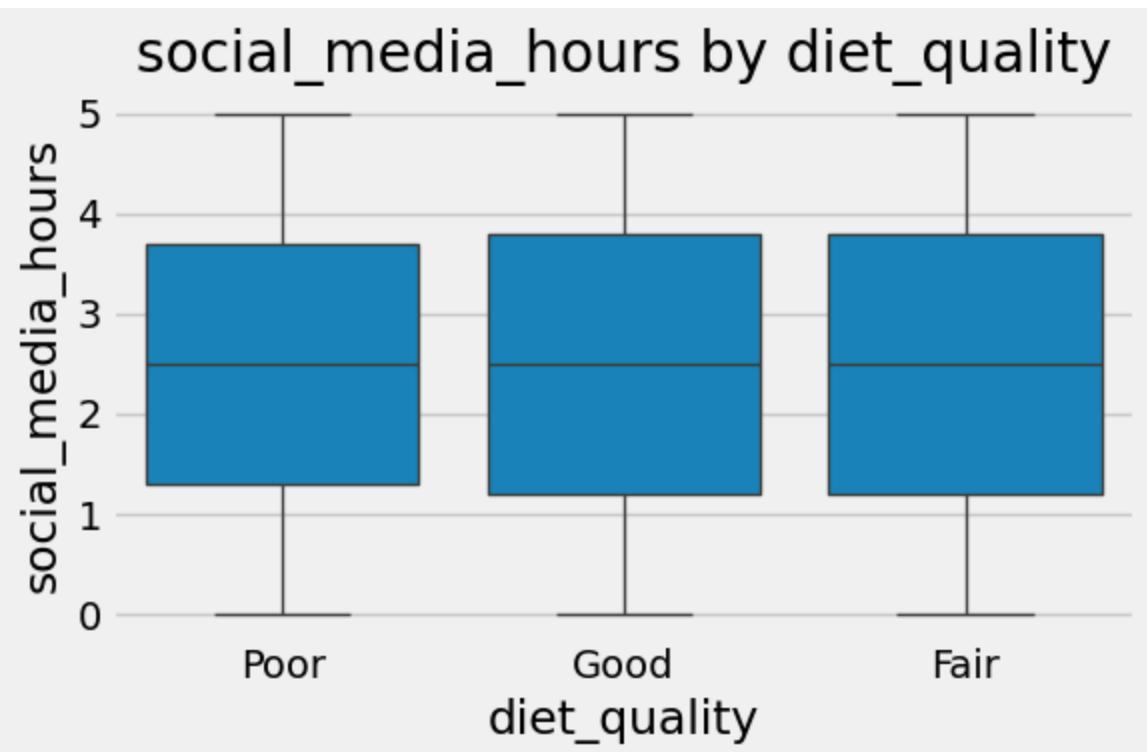
social_media_hours vs exam_score (corr=-0.01)

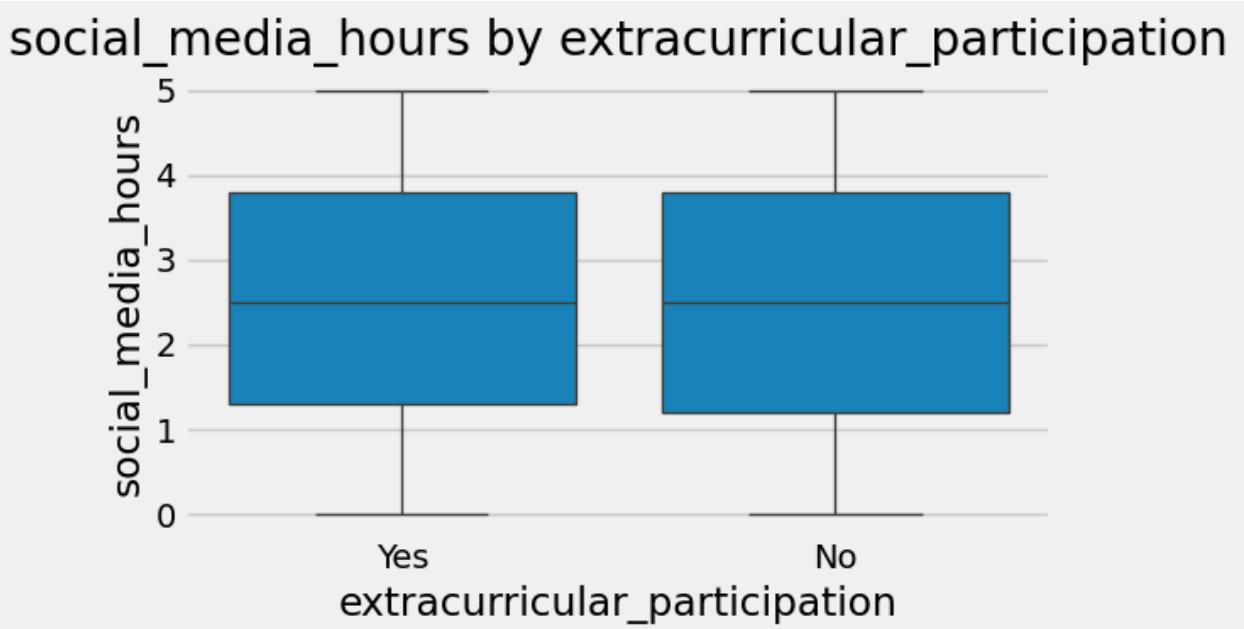
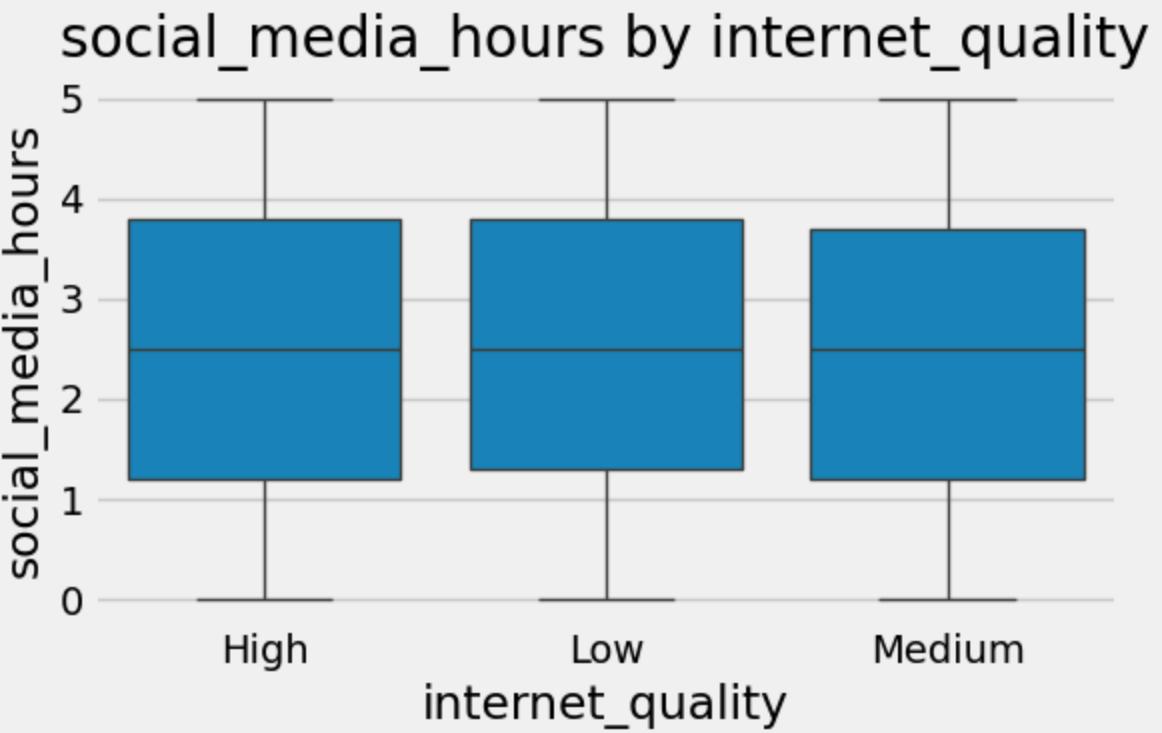


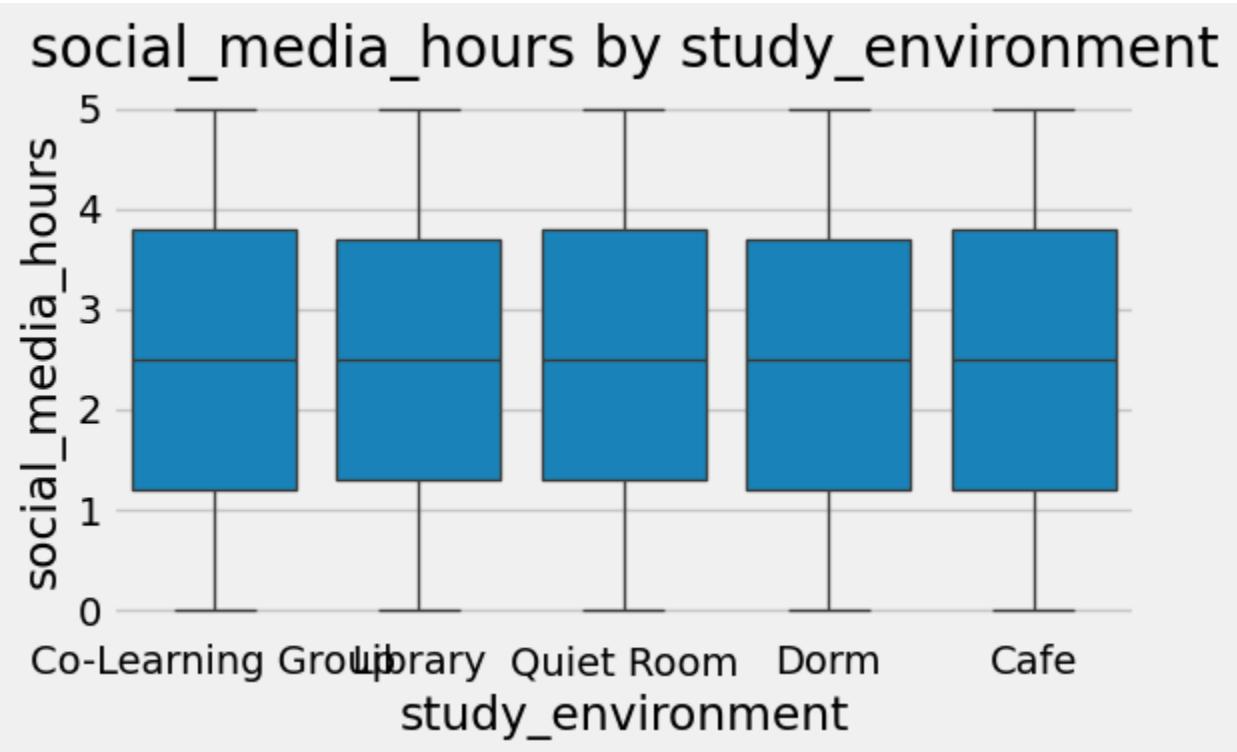
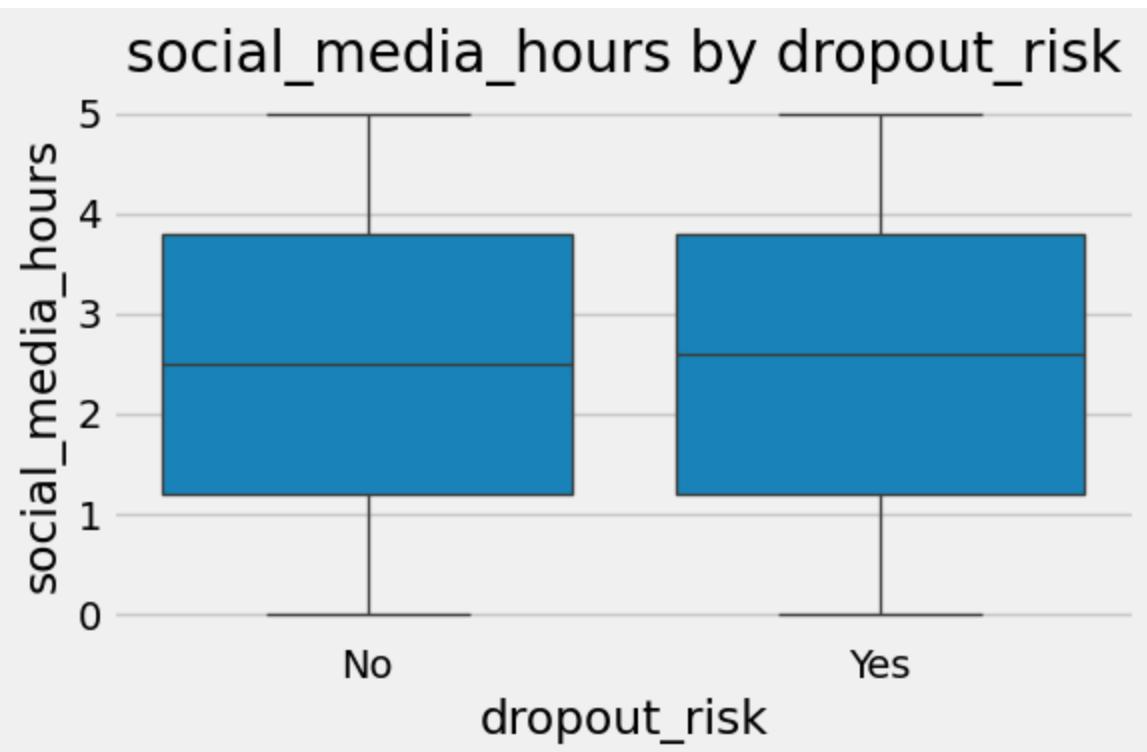
social_media_hours by gender

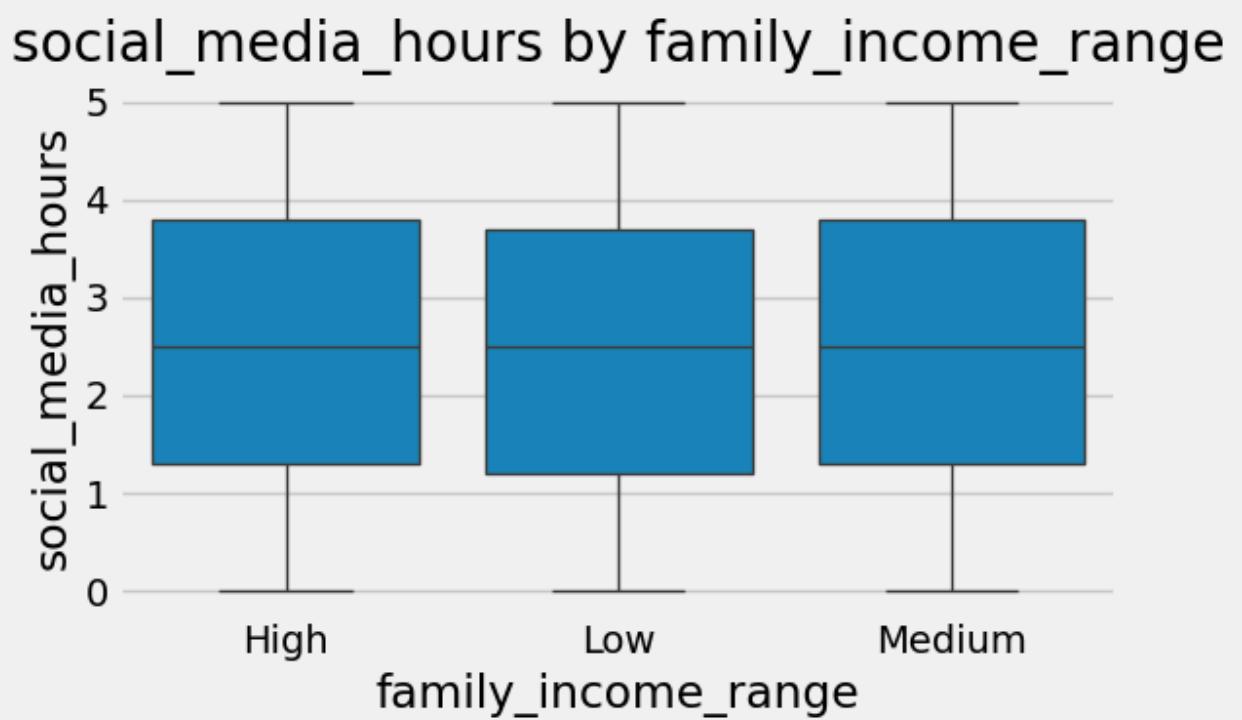
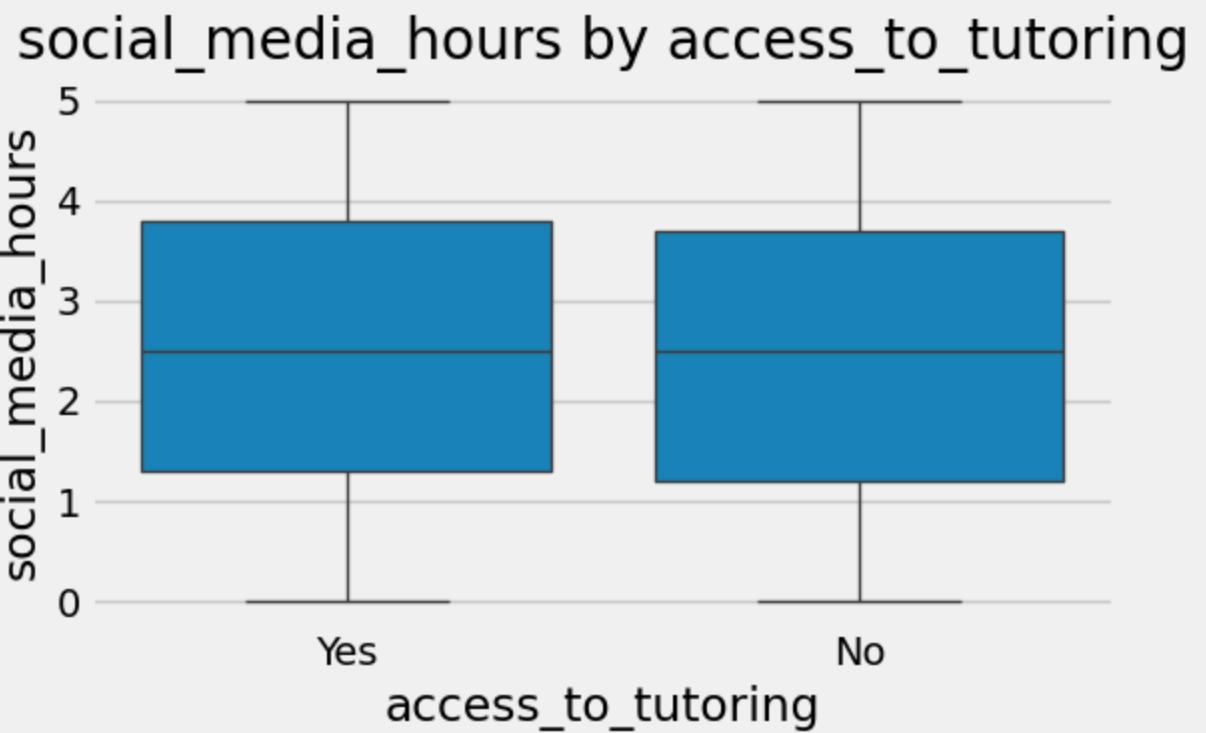


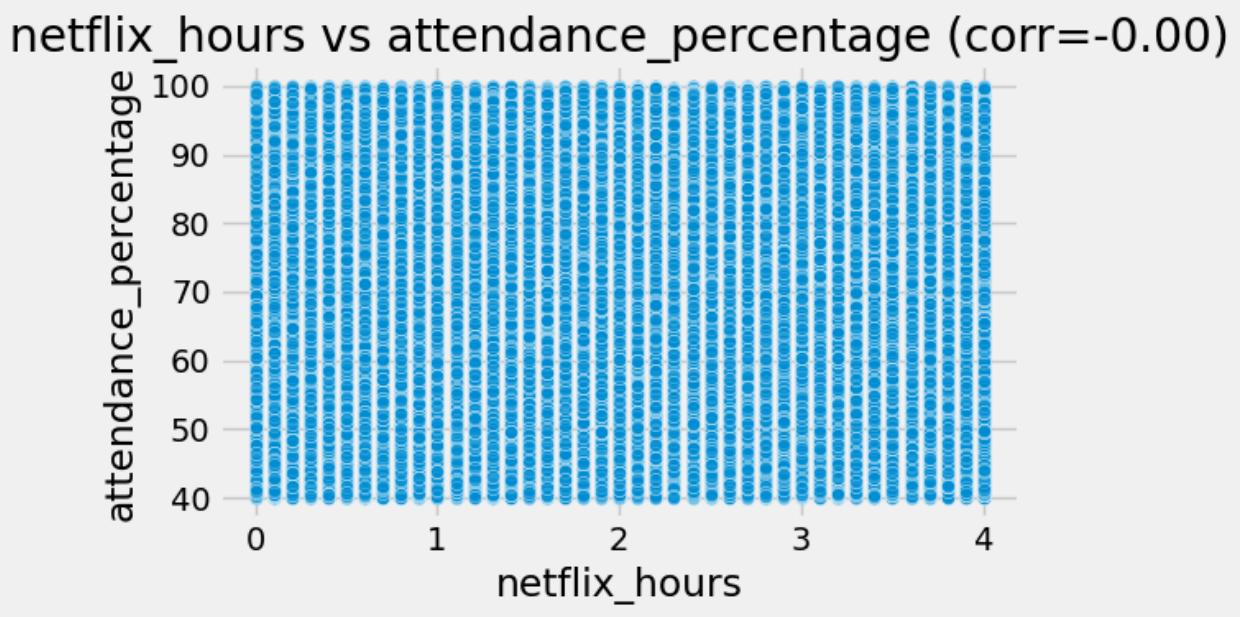
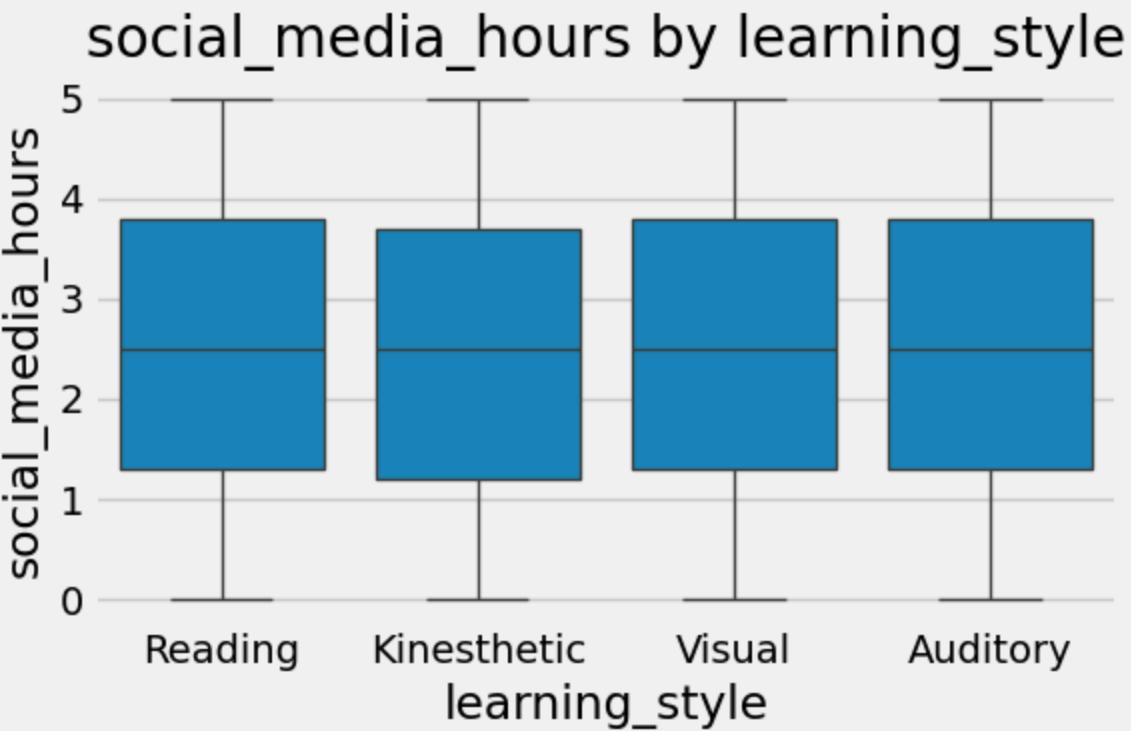




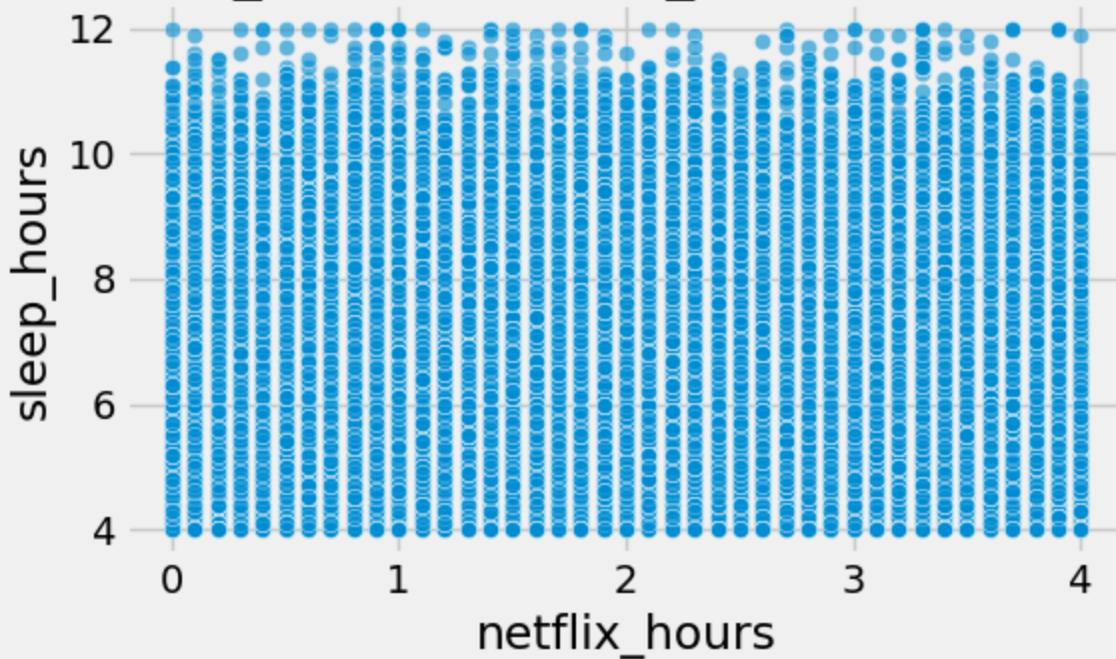




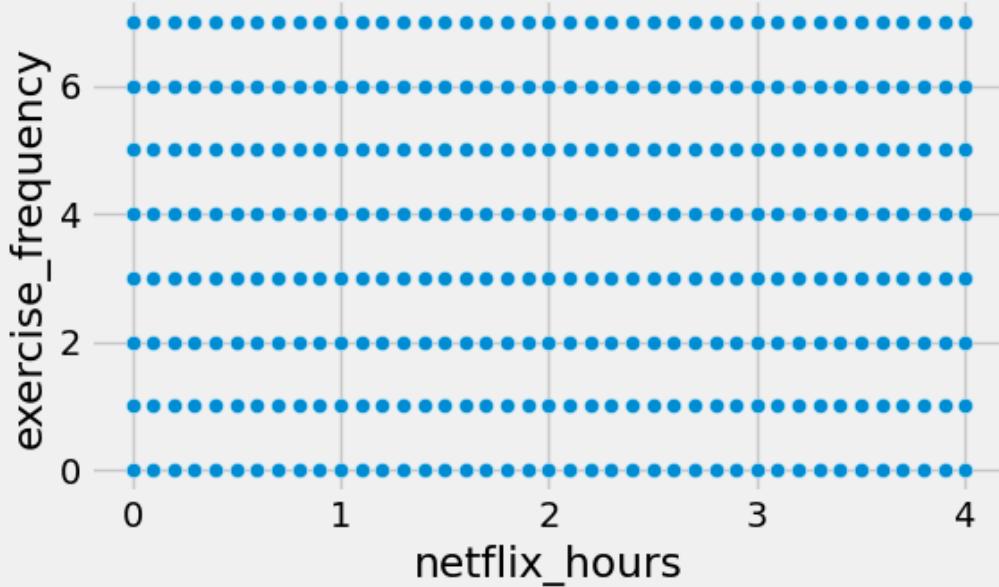




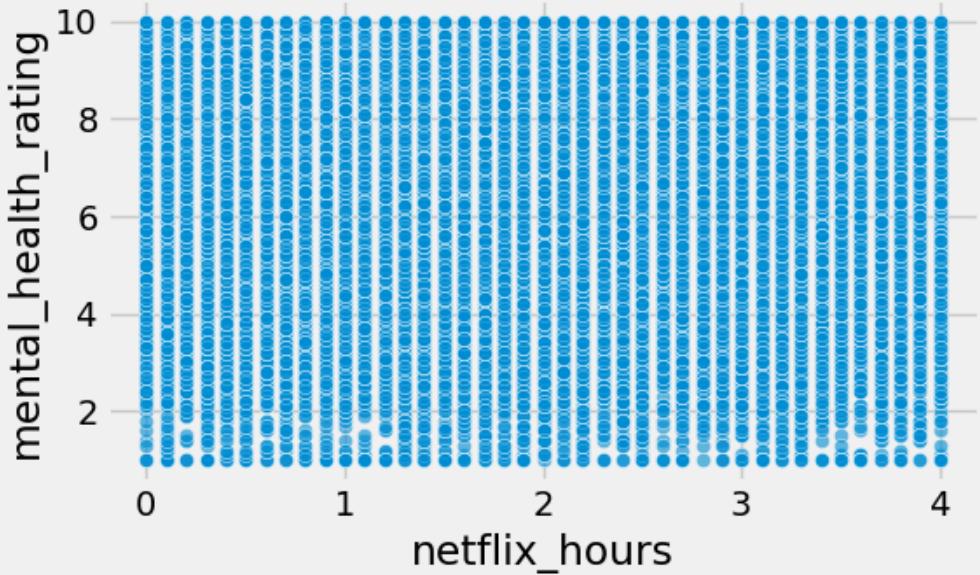
netflix_hours vs sleep_hours (corr=0.00)



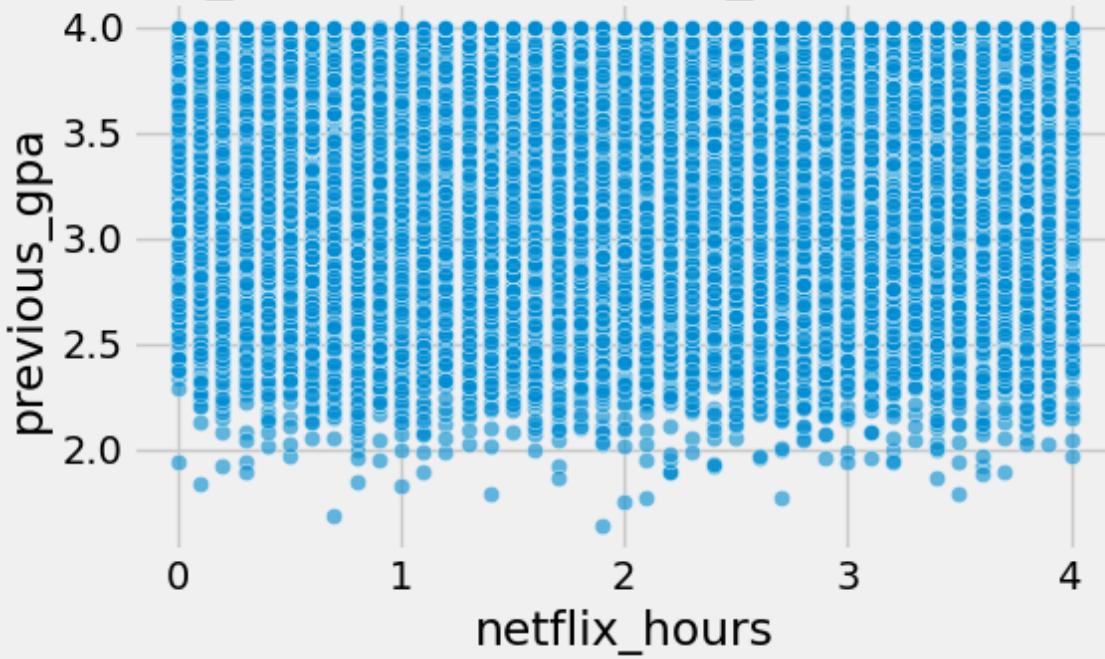
netflix_hours vs exercise_frequency (corr=-0.01)



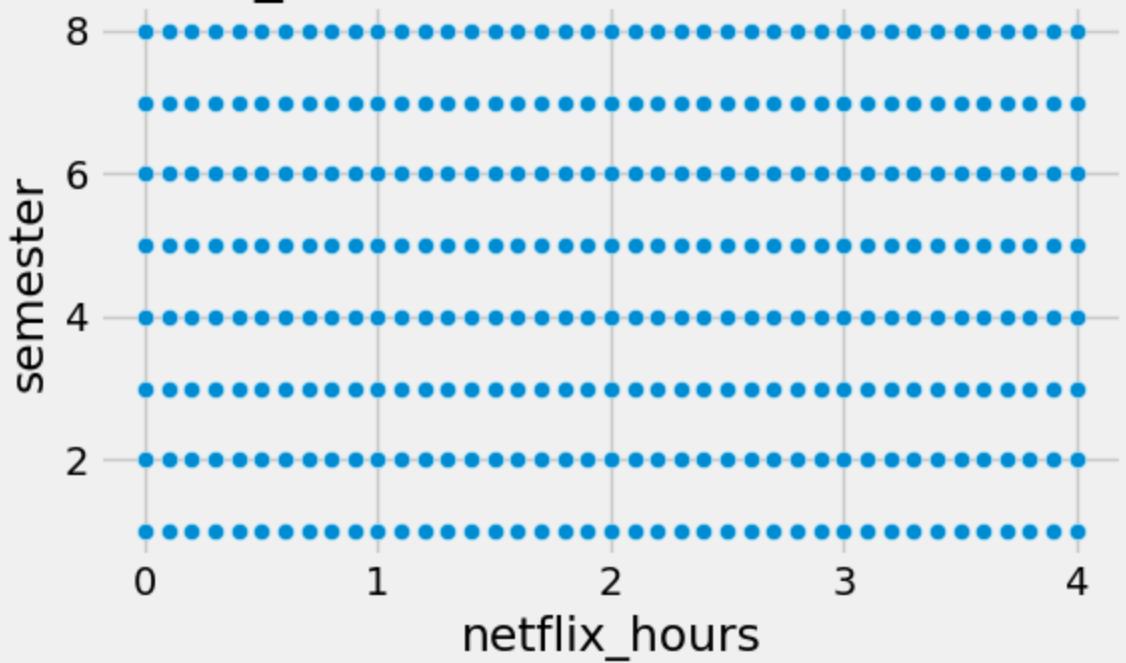
netflix_hours vs mental_health_rating (corr=0.00)



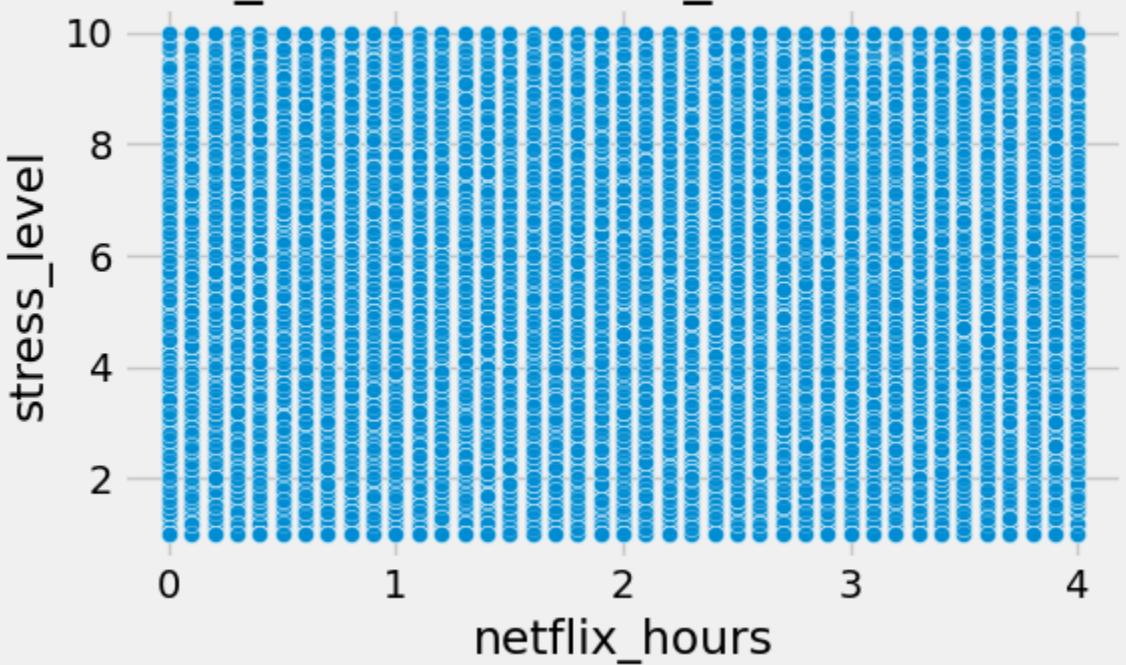
netflix_hours vs previous_gpa (corr=-0.00)



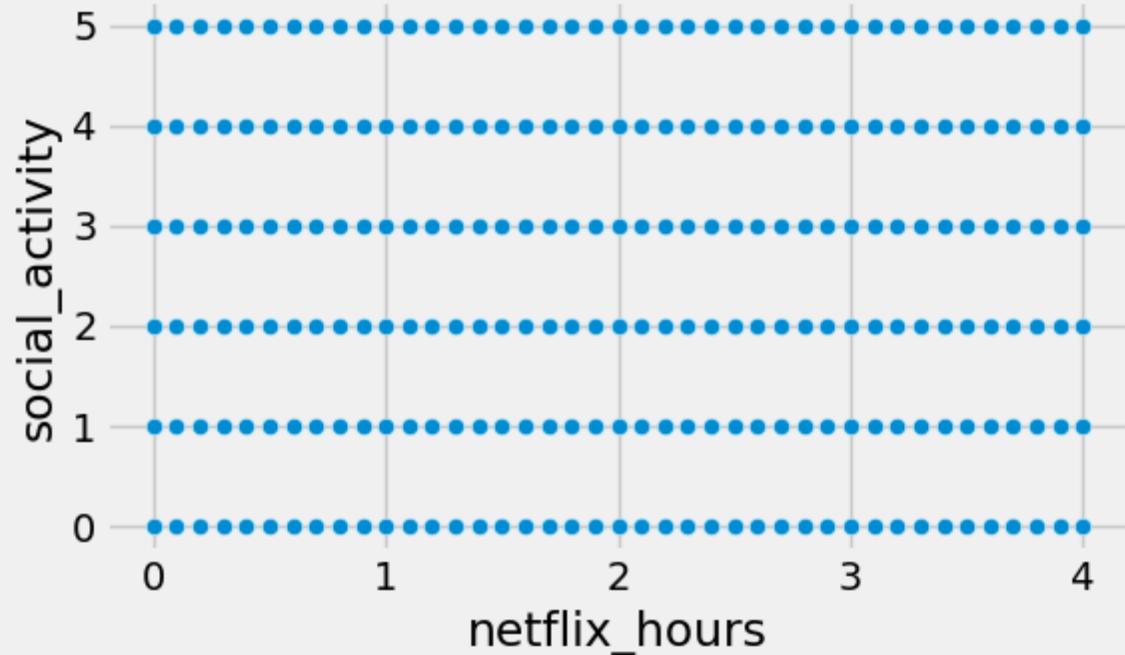
netflix_hours vs semester (corr=0.00)



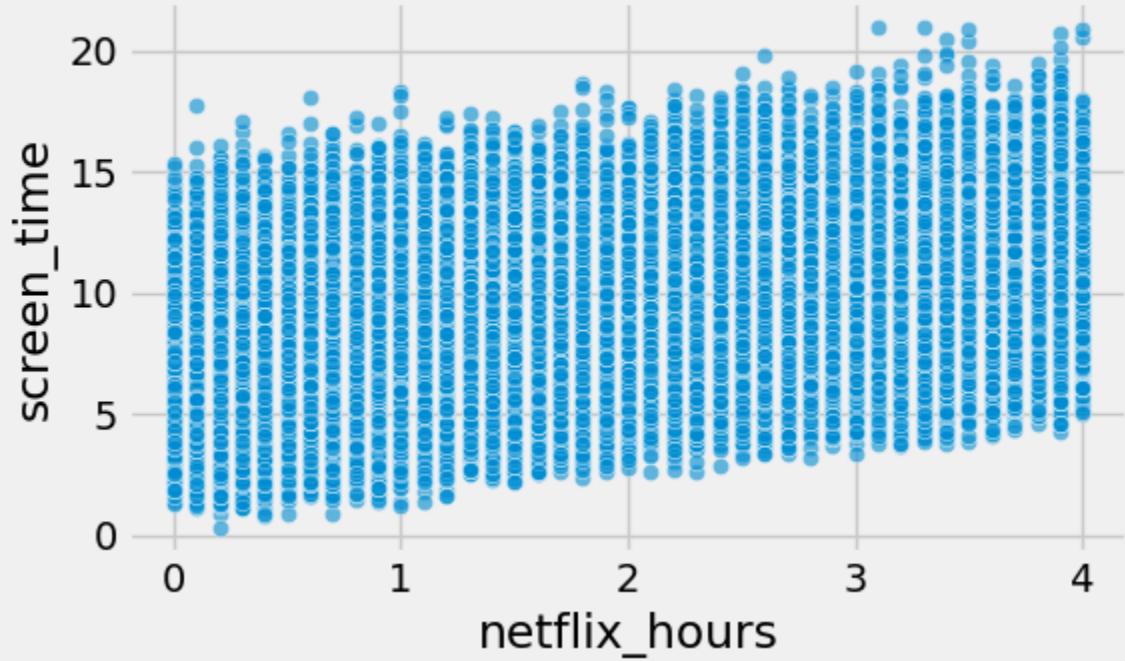
netflix_hours vs stress_level (corr=0.00)



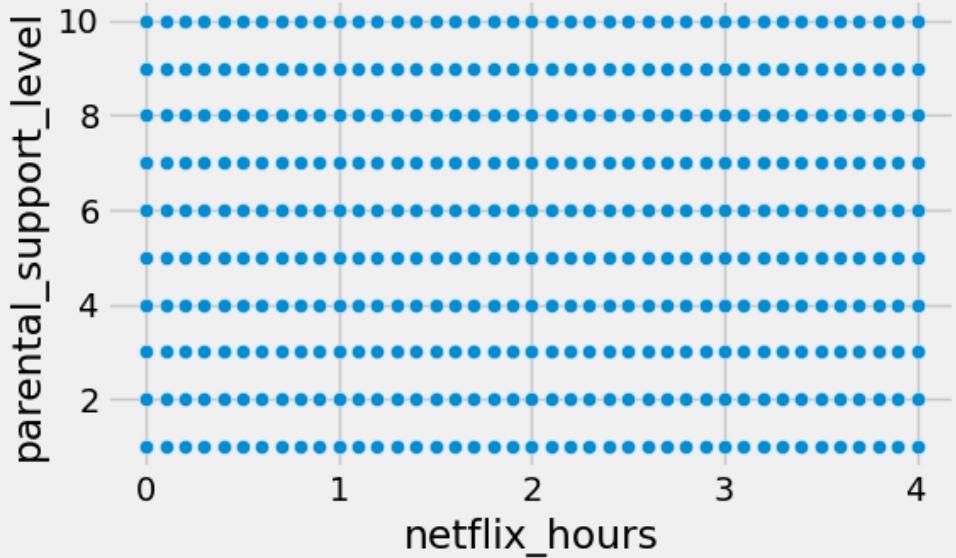
netflix_hours vs social_activity (corr=-0.00)



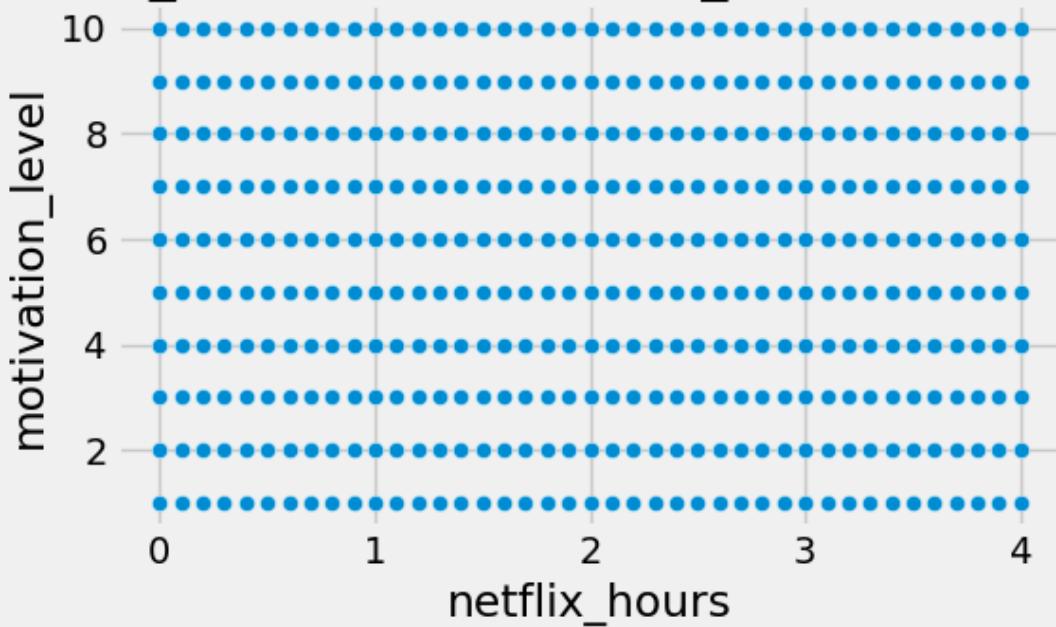
netflix_hours vs screen_time (corr=0.41)



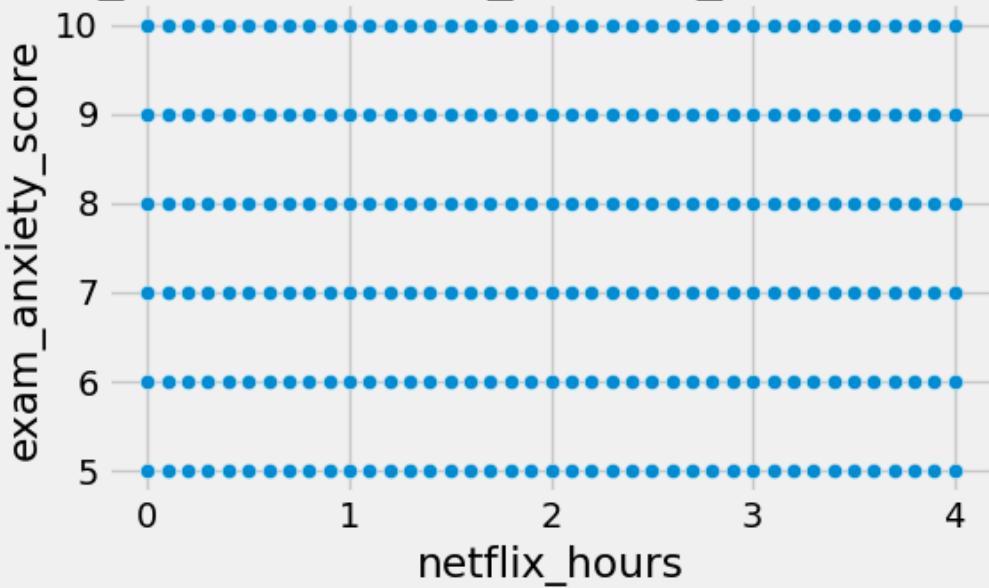
netflix_hours vs parental_support_level (corr=0.00)



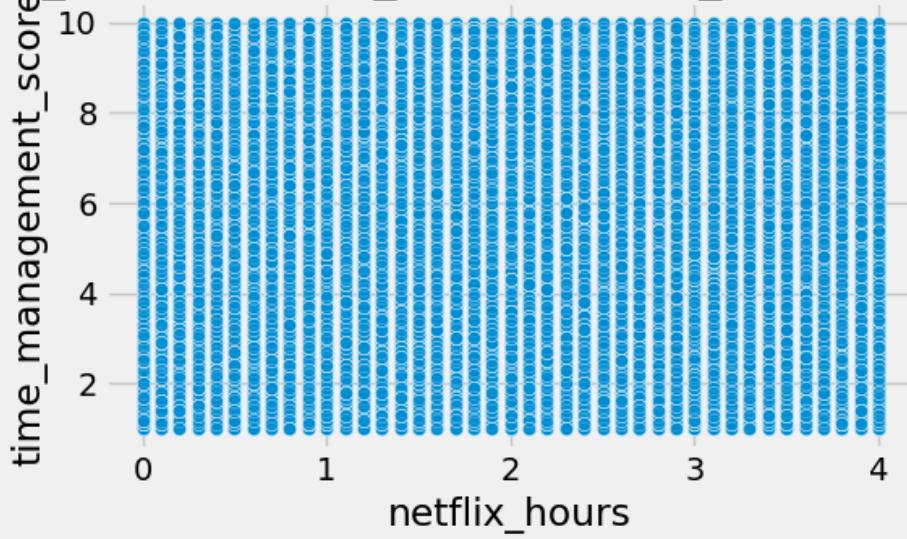
netflix_hours vs motivation_level (corr=-0.00)



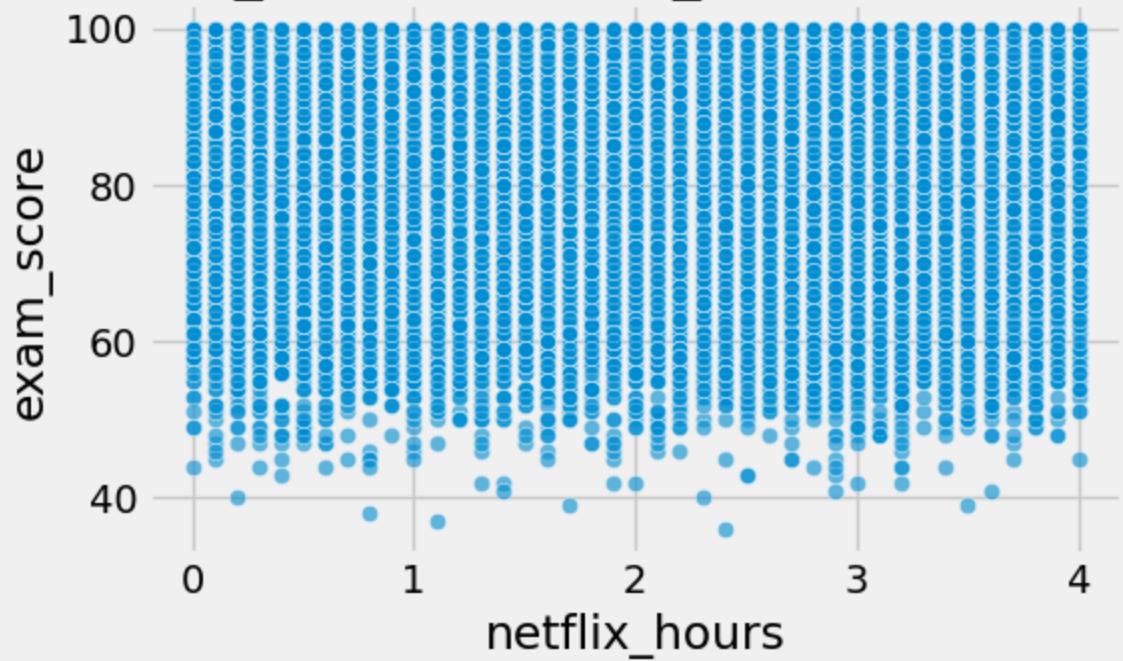
netflix_hours vs exam_anxiety_score (corr=0.00)



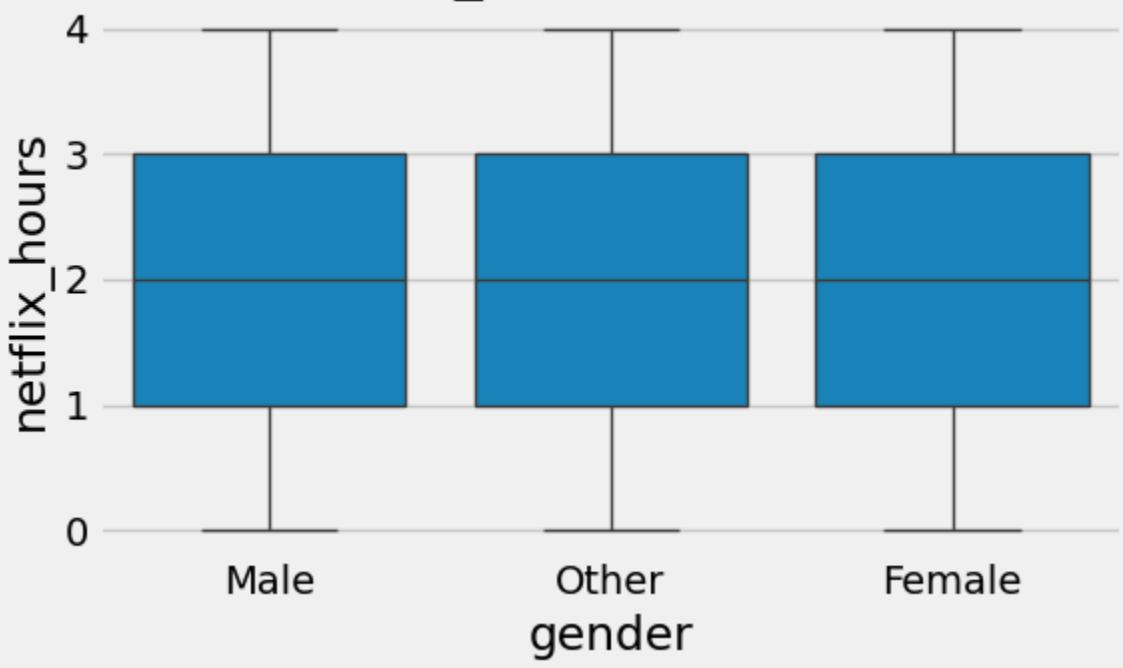
netflix_hours vs time_management_score (corr=0.01)

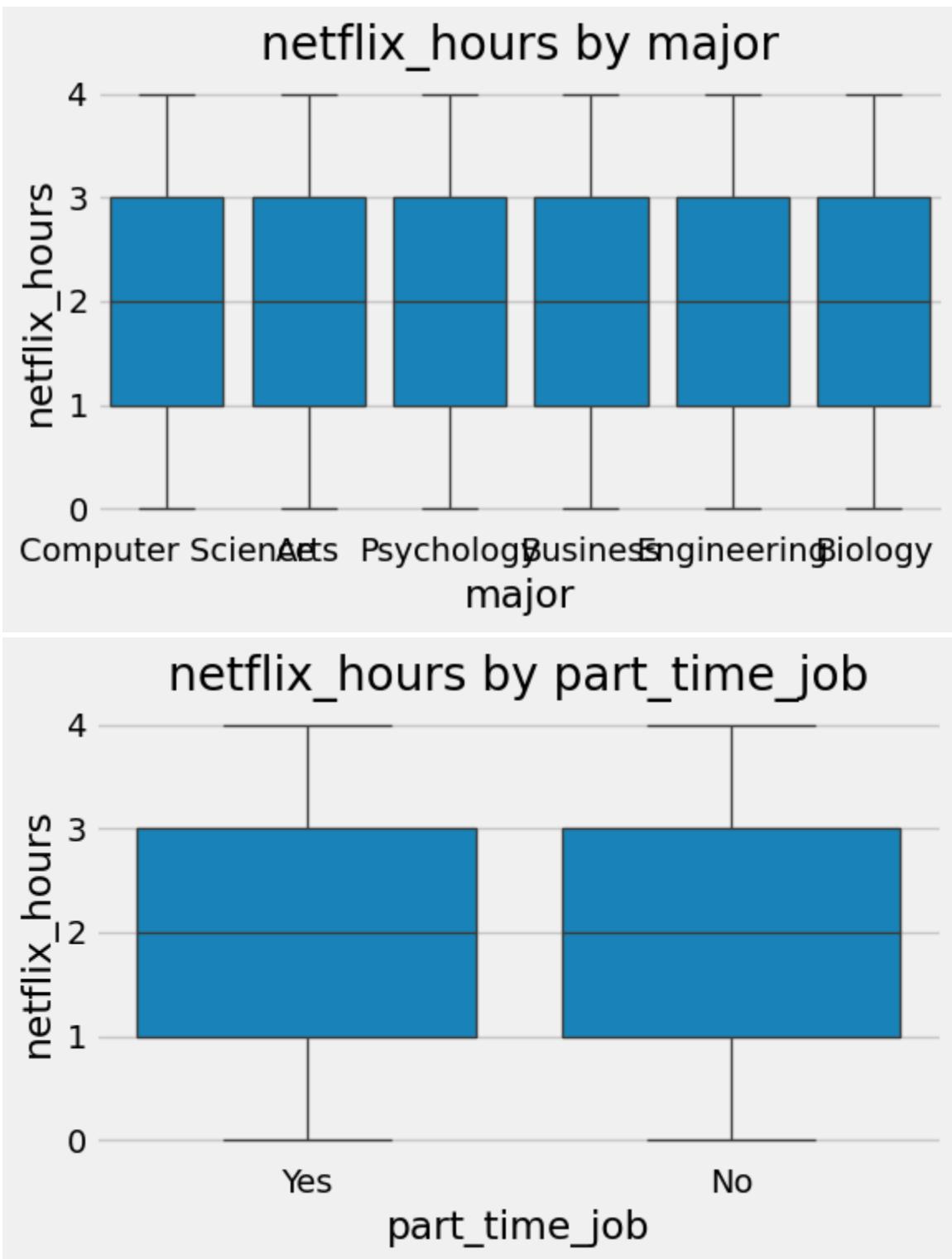


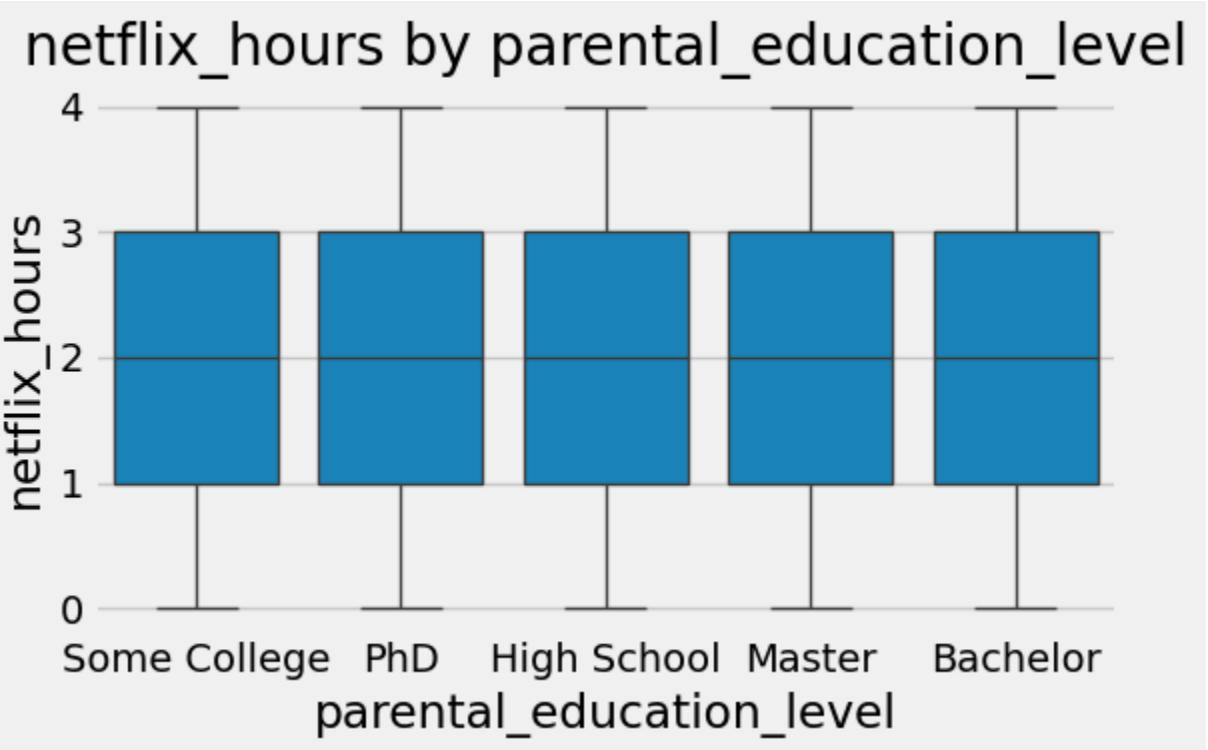
`netflix_hours` vs `exam_score` (corr=-0.00)

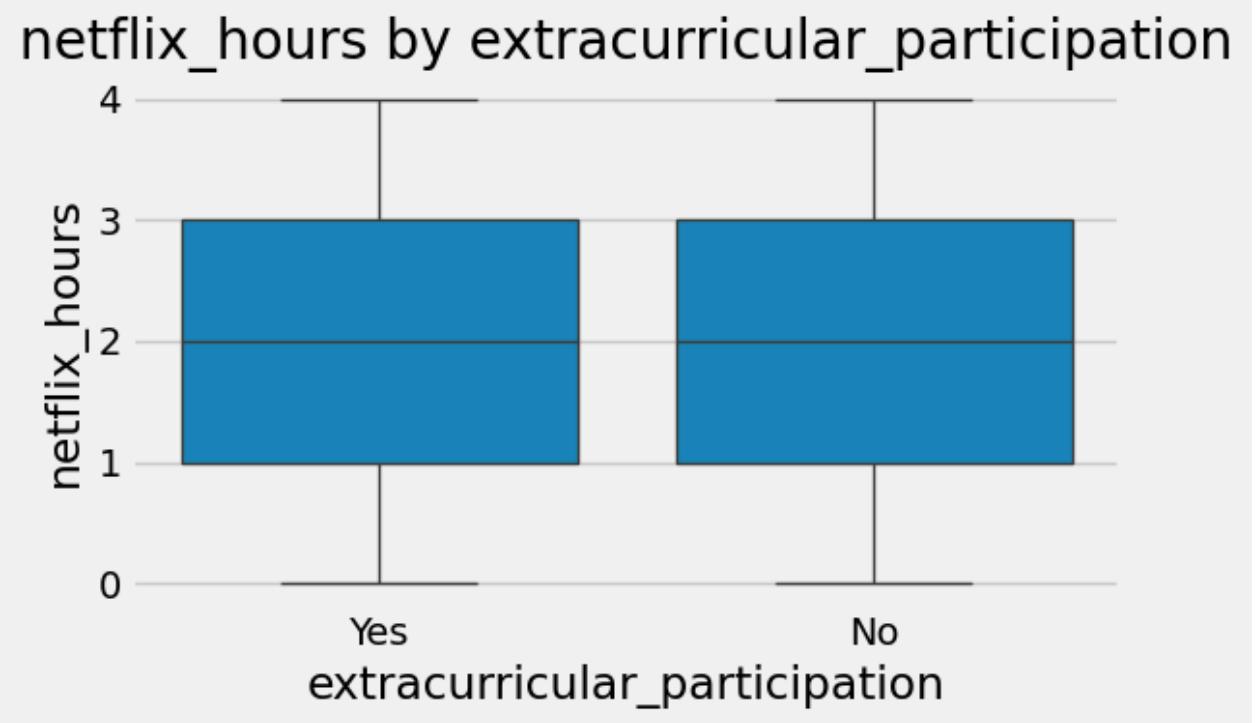
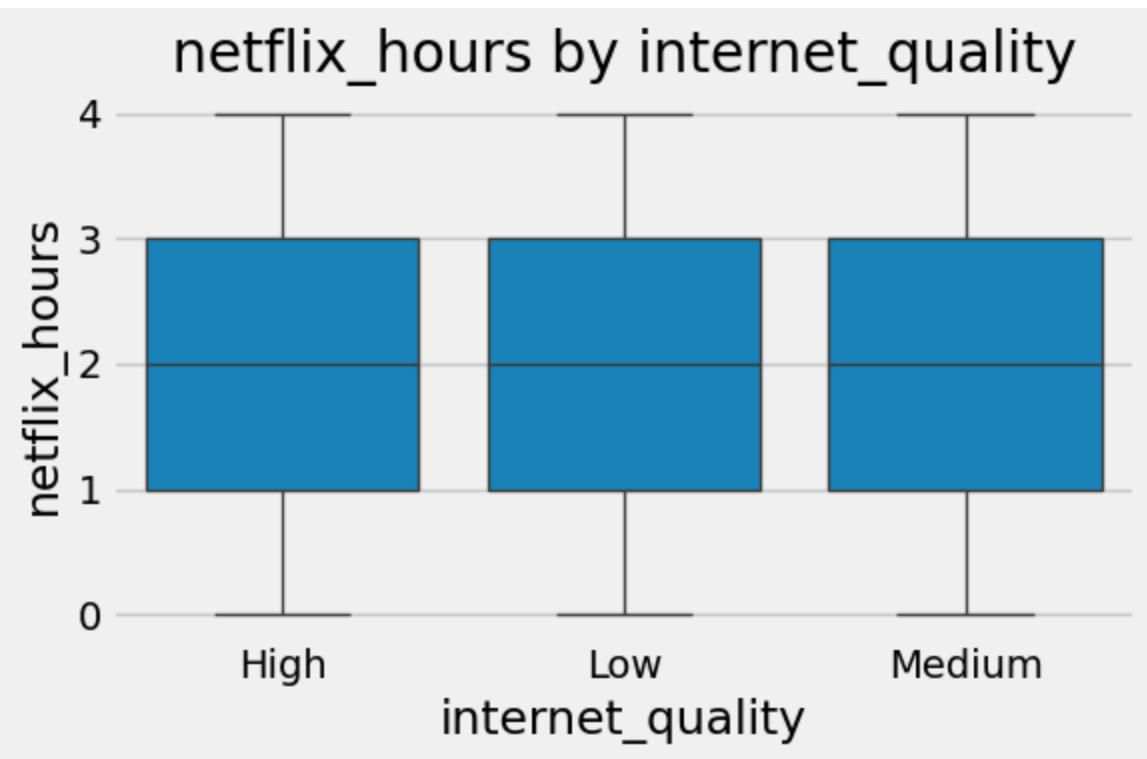


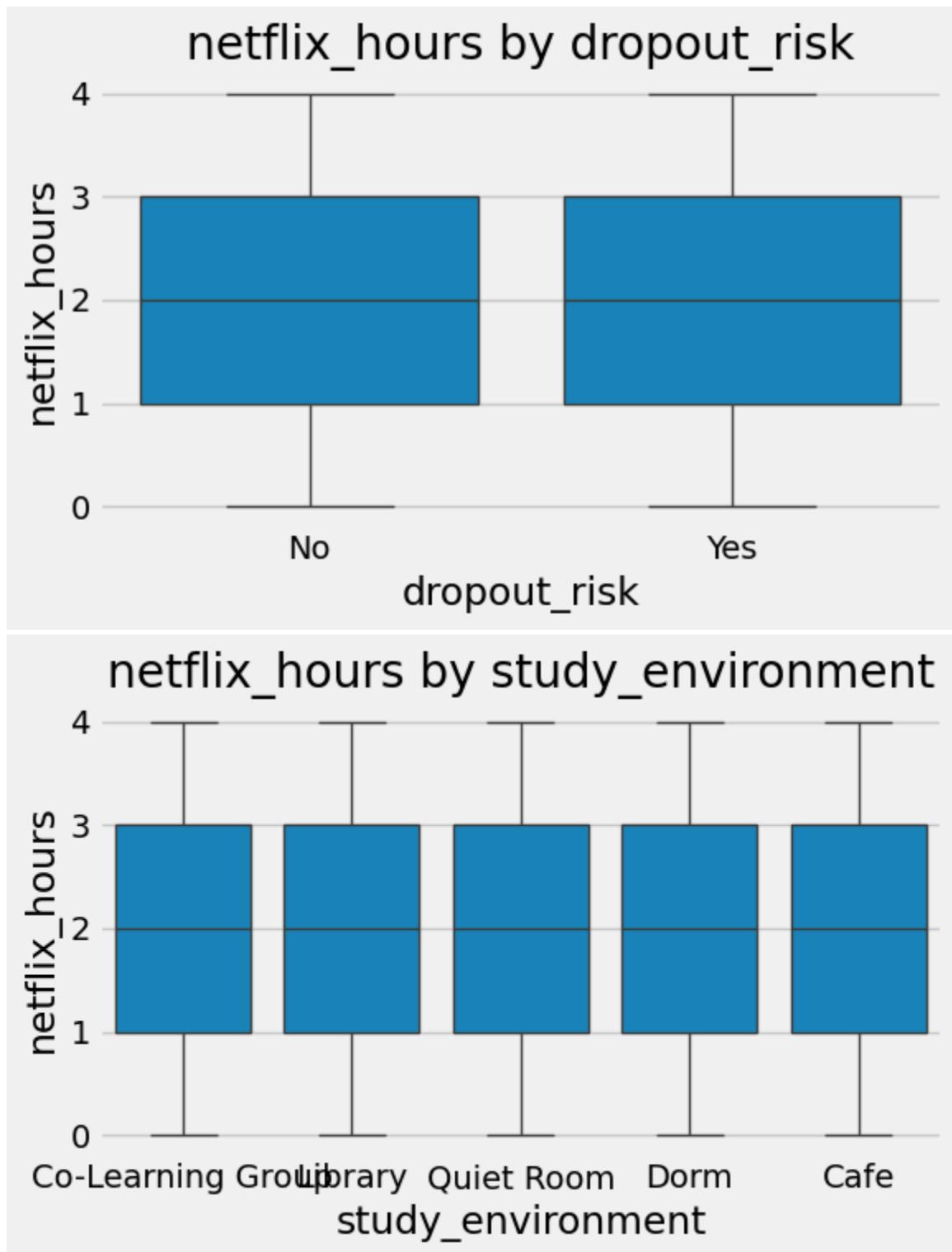
`netflix_hours` by gender

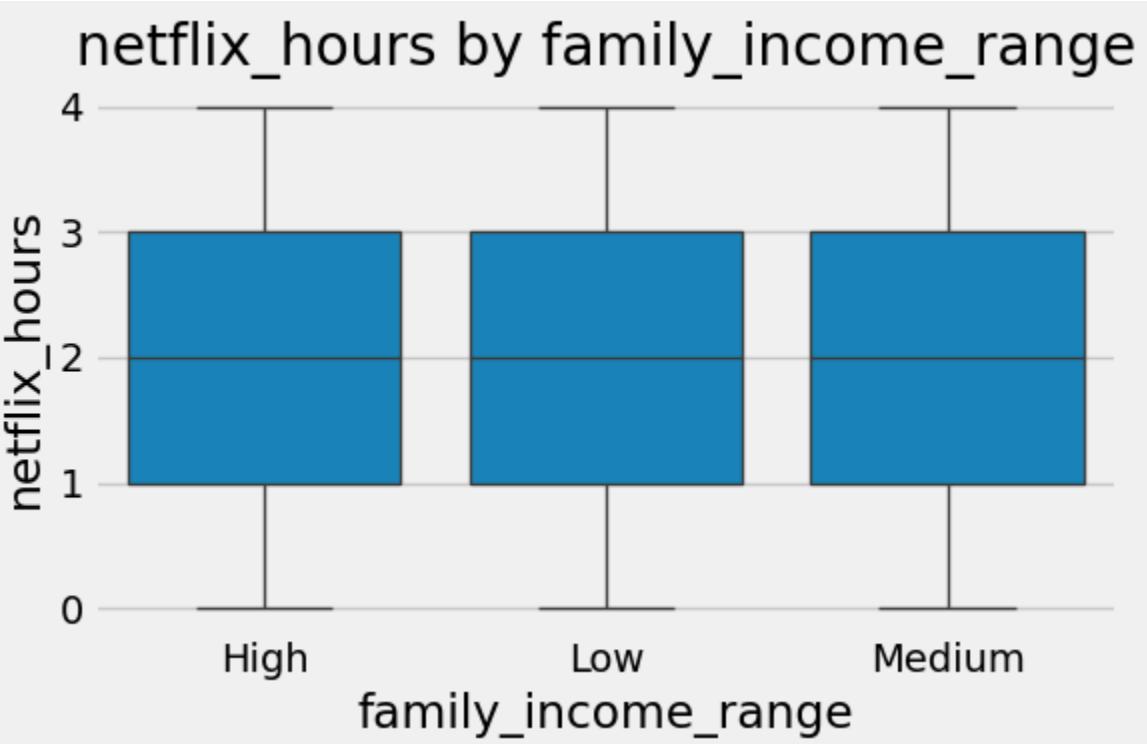
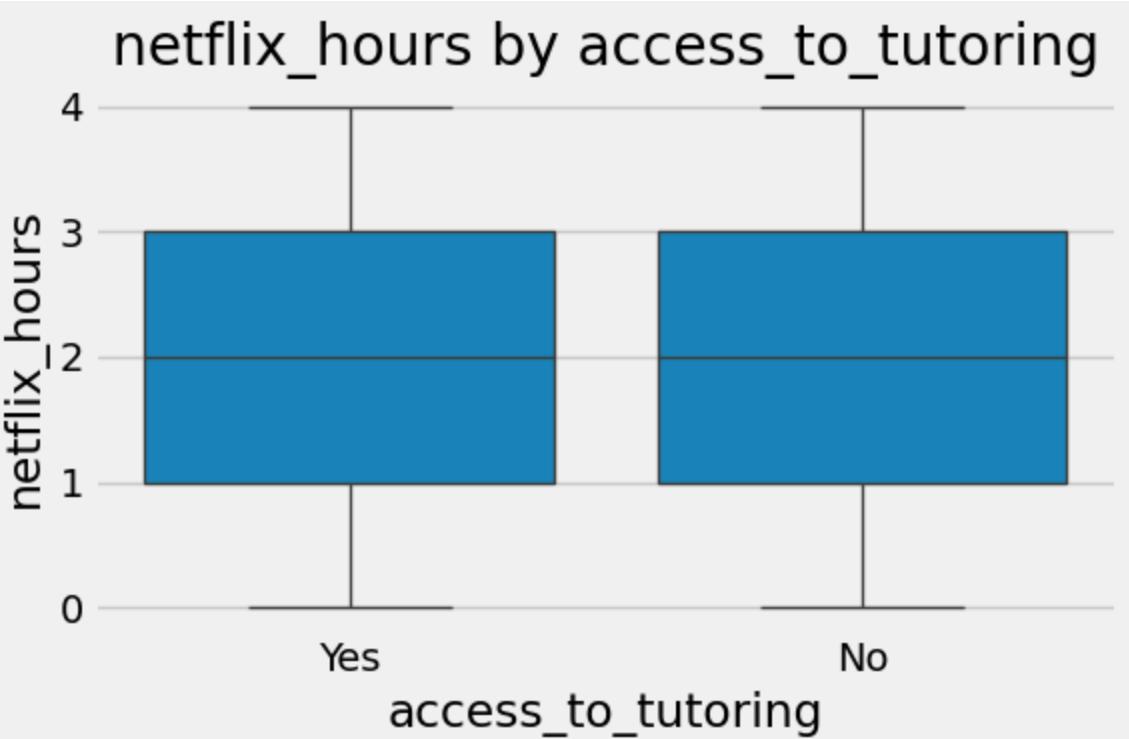


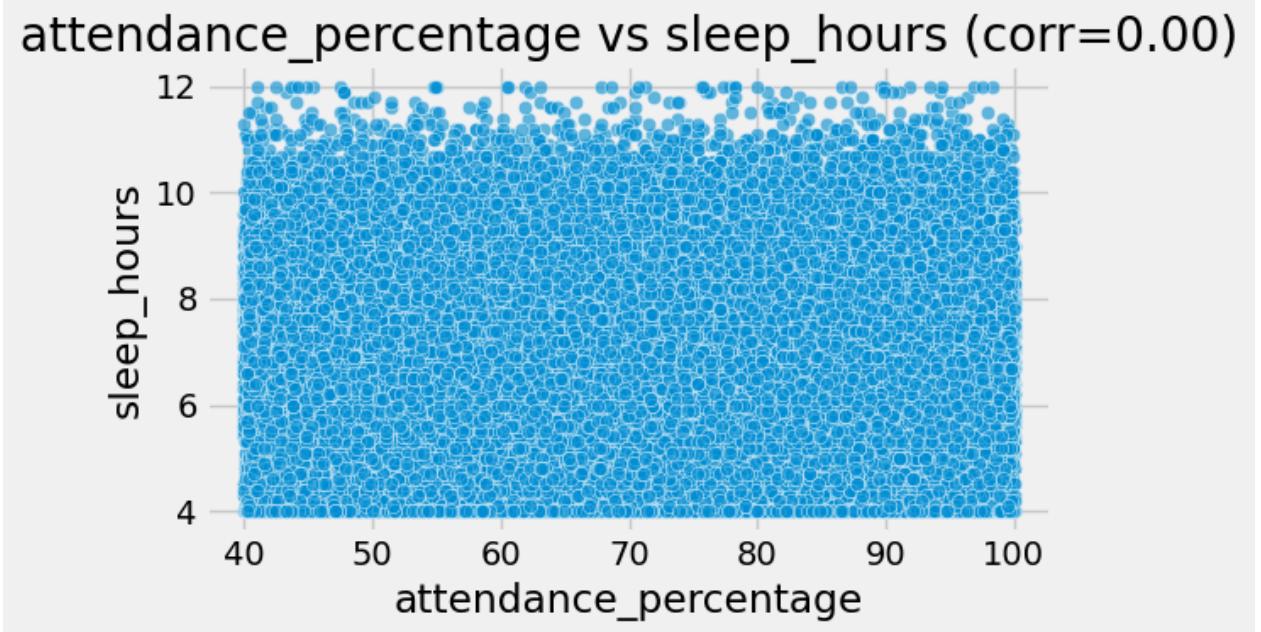
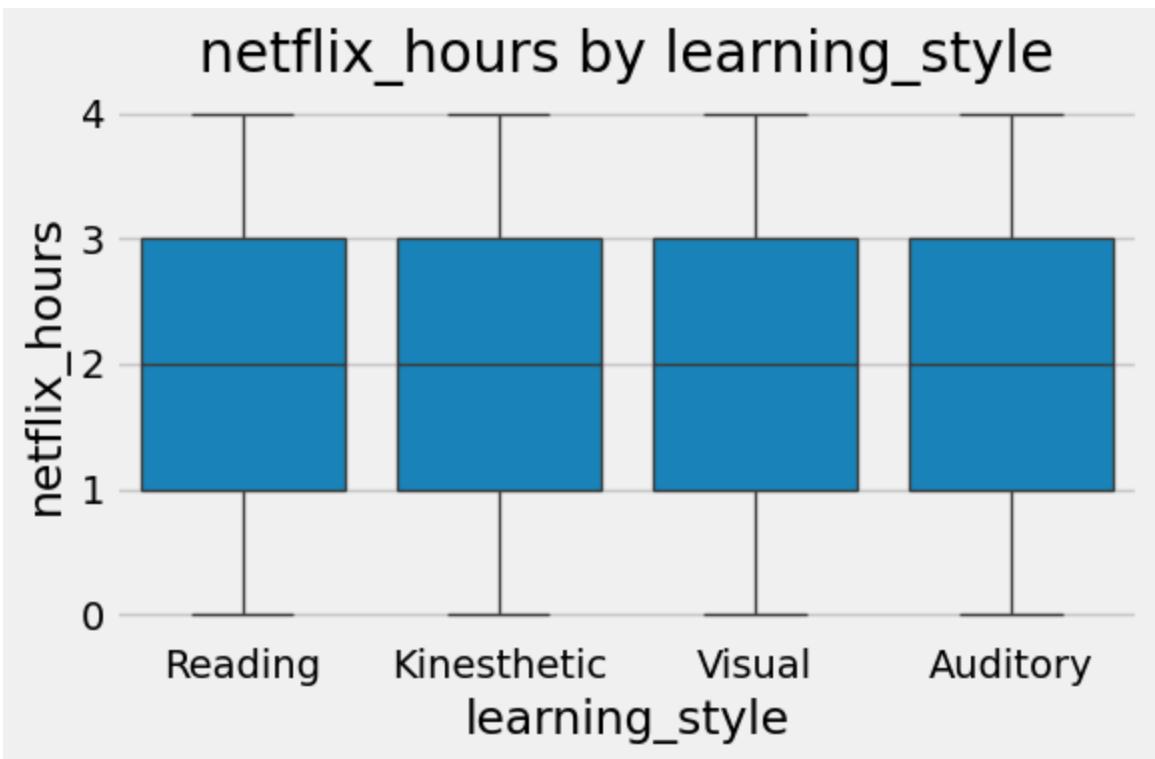




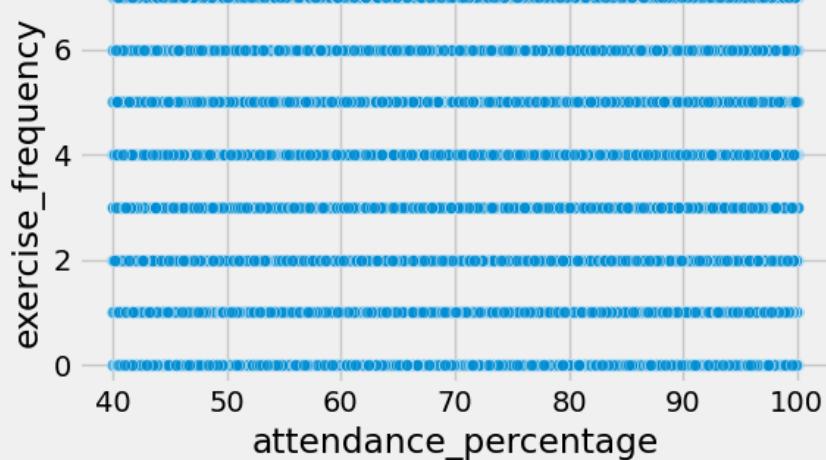




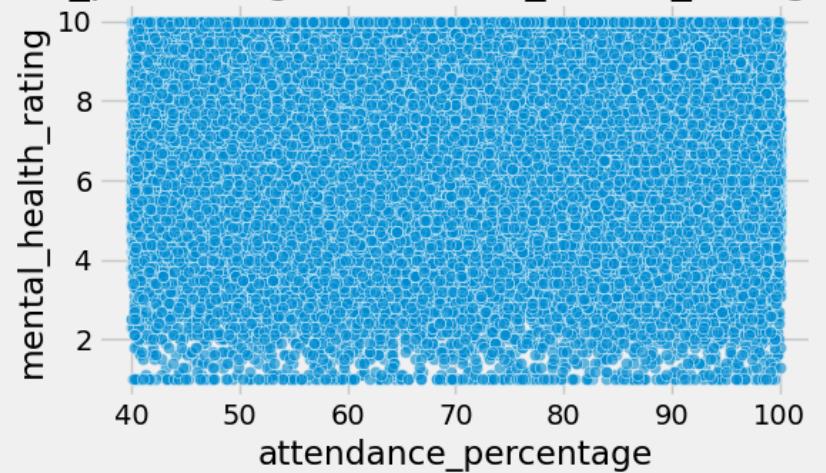




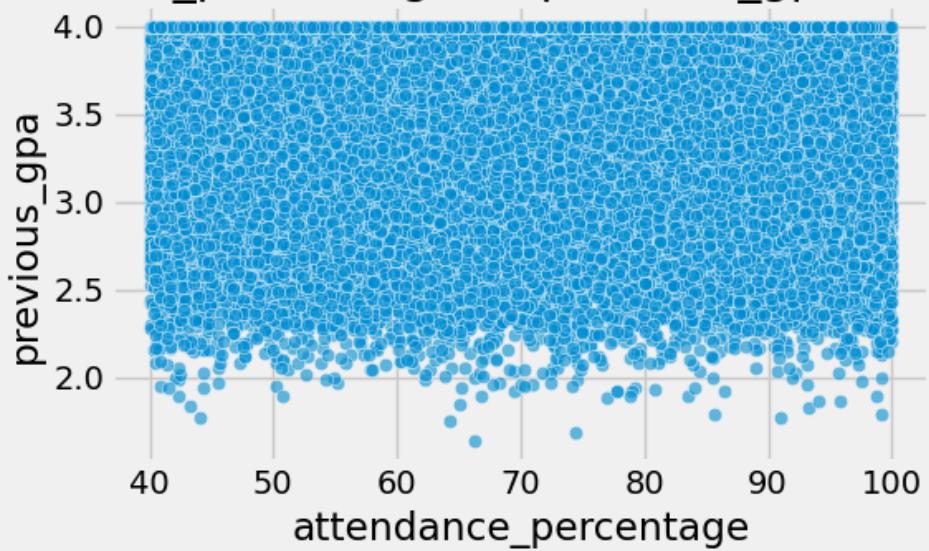
attendance_percentage vs exercise_frequency (corr=-0.00)



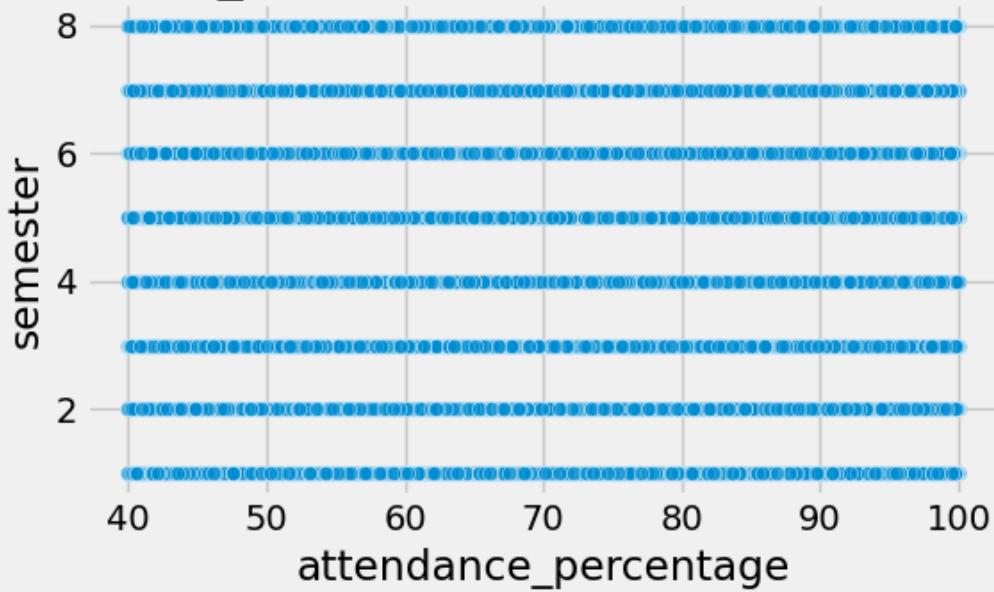
attendance_percentage vs mental_health_rating (corr=-0.00)



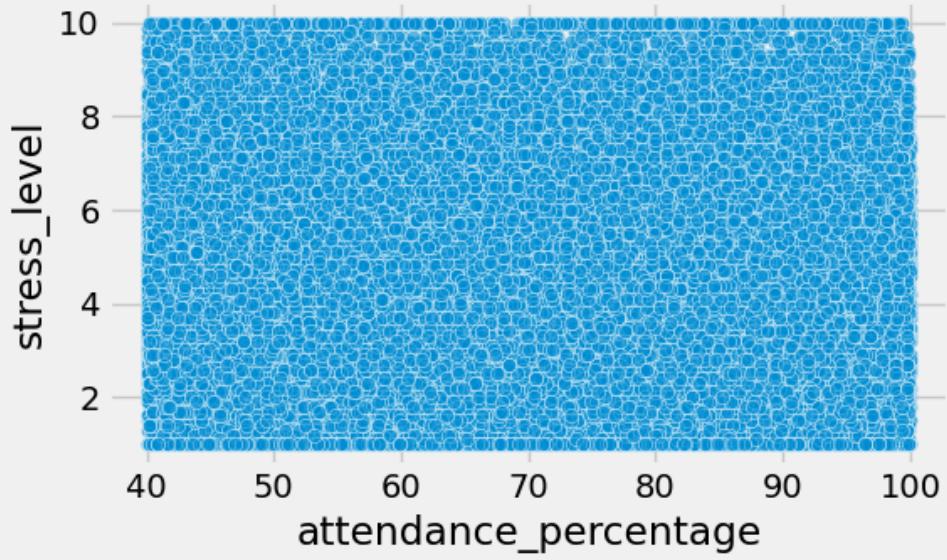
attendance_percentage vs previous_gpa (corr=0.00)



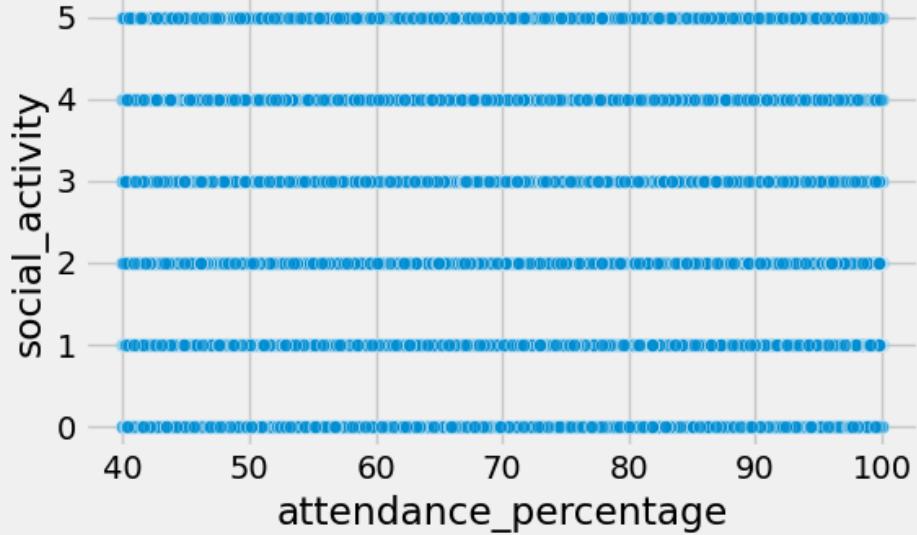
attendance_percentage vs semester (corr=0.00)



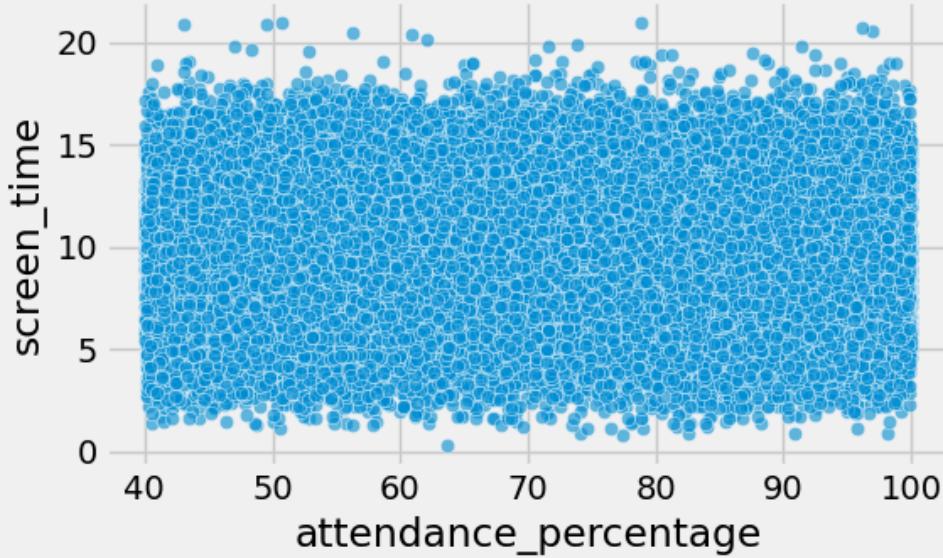
attendance_percentage vs stress_level (corr=-0.00)



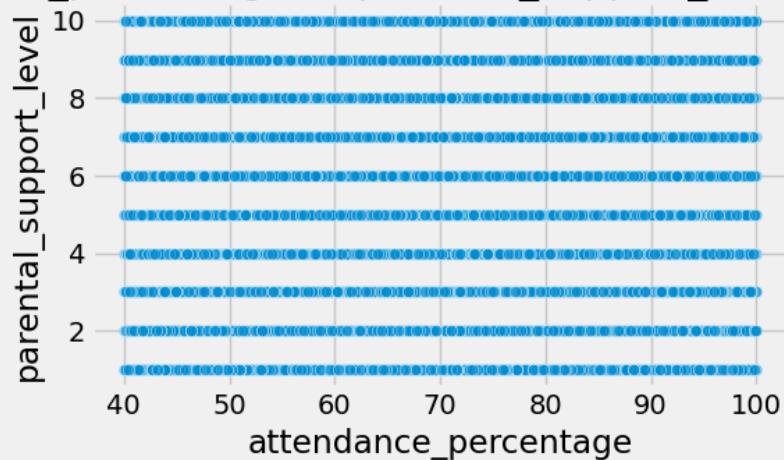
attendance_percentage vs social_activity (corr=0.00)



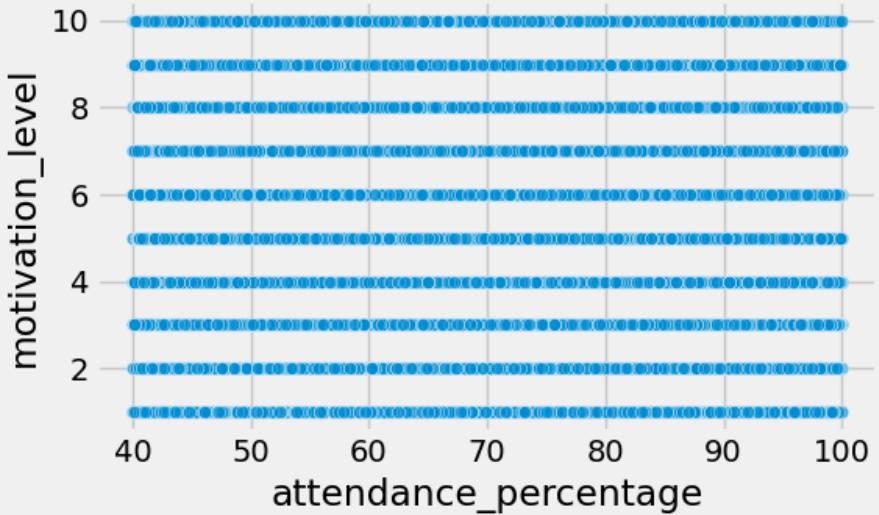
attendance_percentage vs screen_time (corr=0.00)



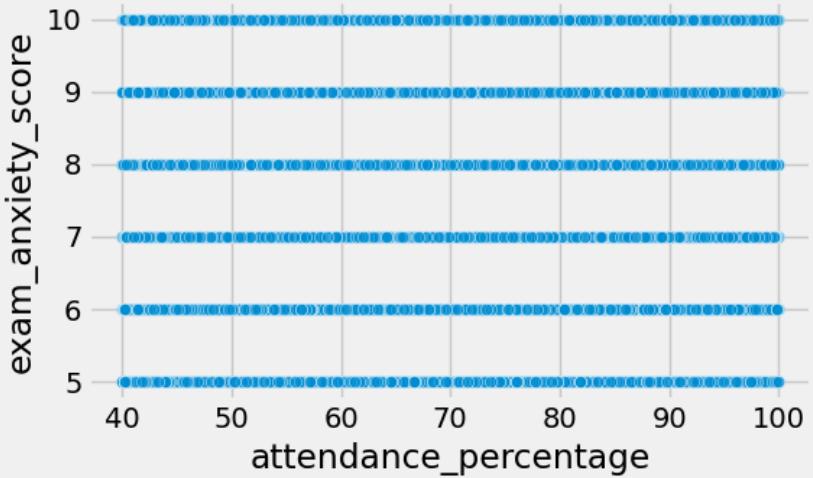
attendance_percentage vs parental_support_level (corr=-0.00)



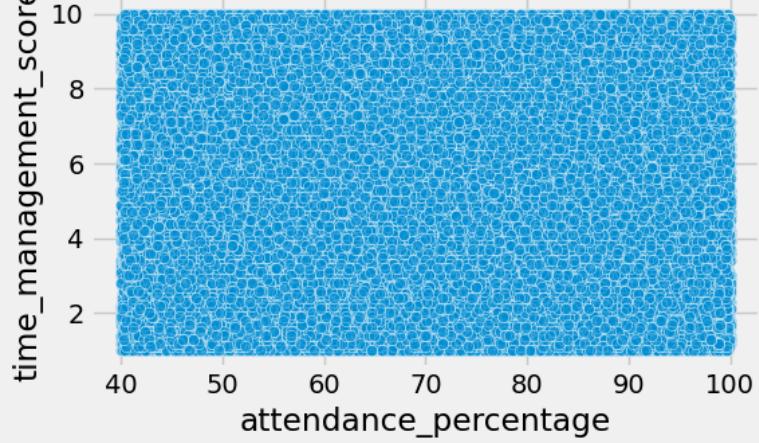
attendance_percentage vs motivation_level (corr=0.00)



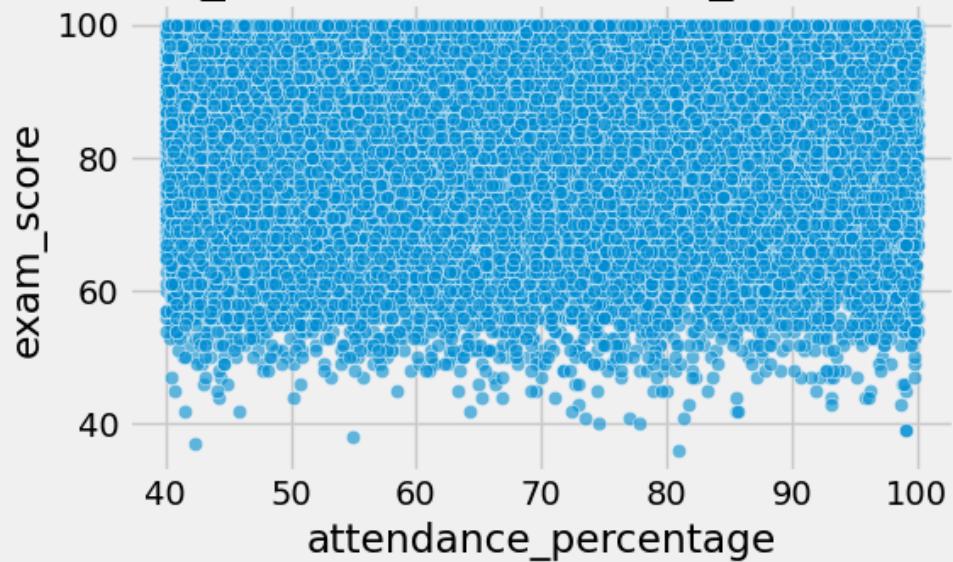
attendance_percentage vs exam_anxiety_score (corr=-0.00)



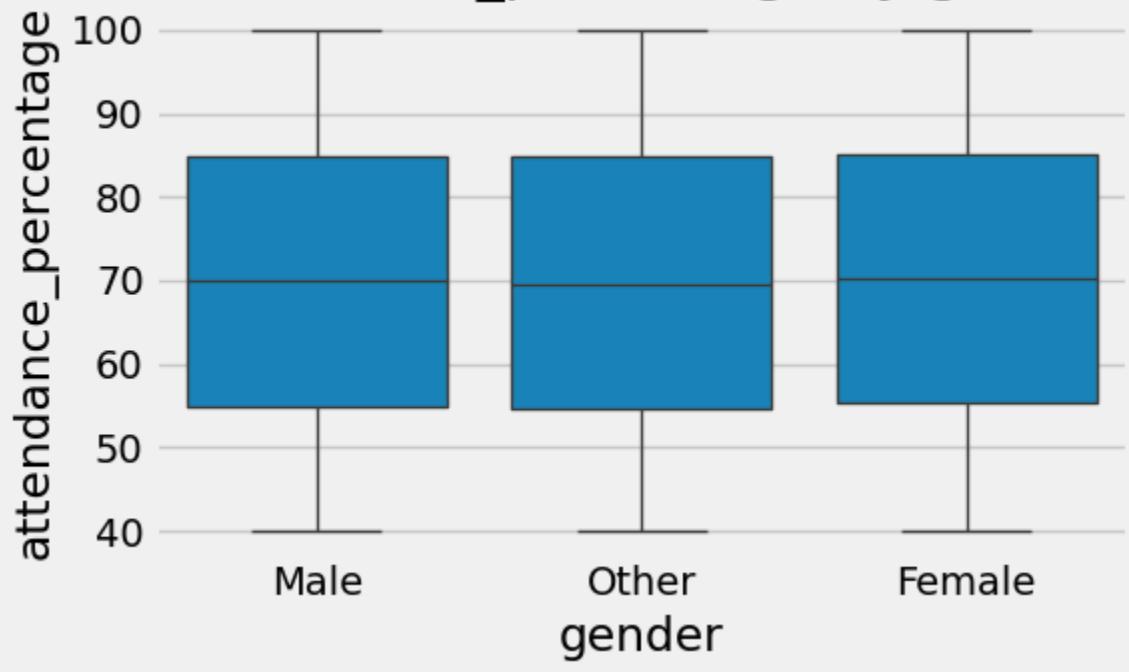
attendance_percentage vs time_management_score (corr=-0.00)

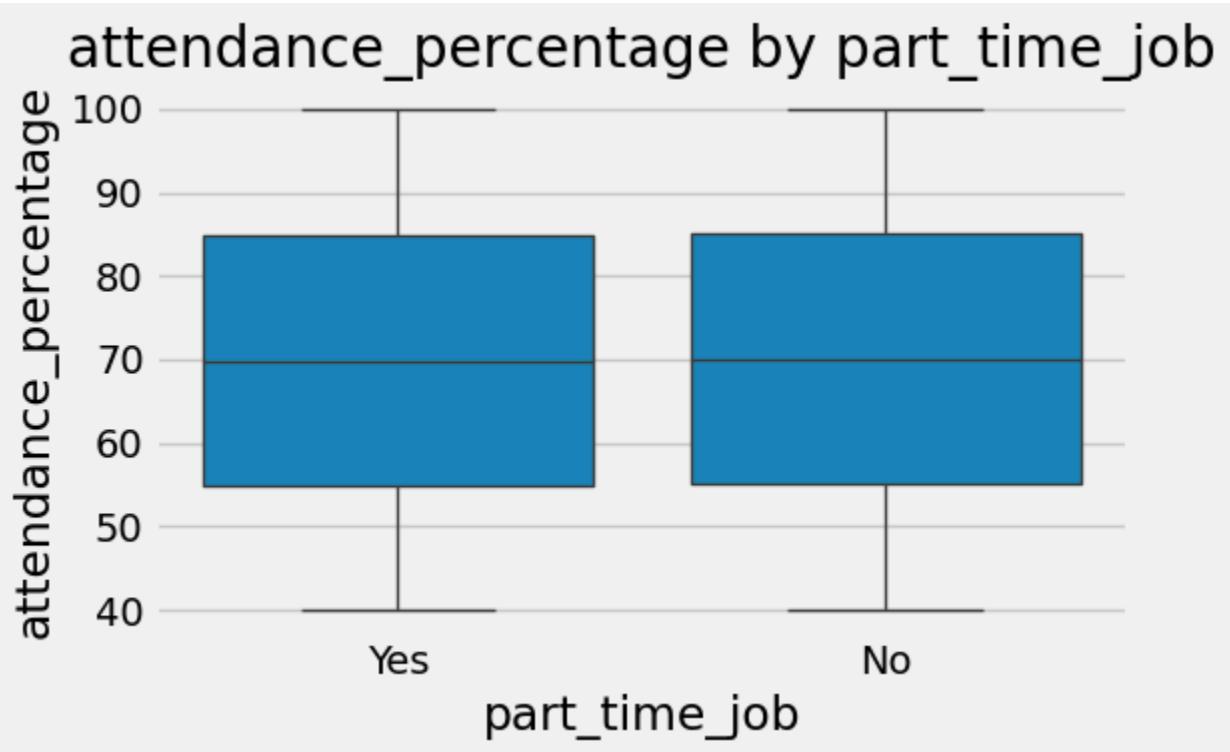
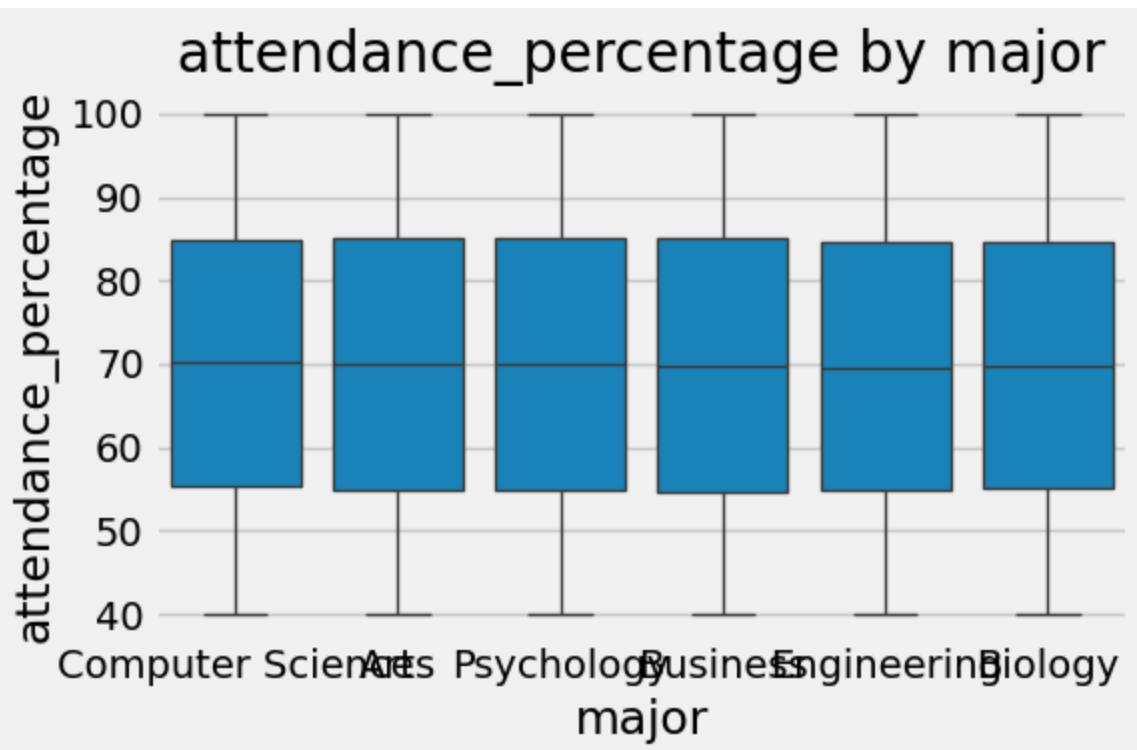


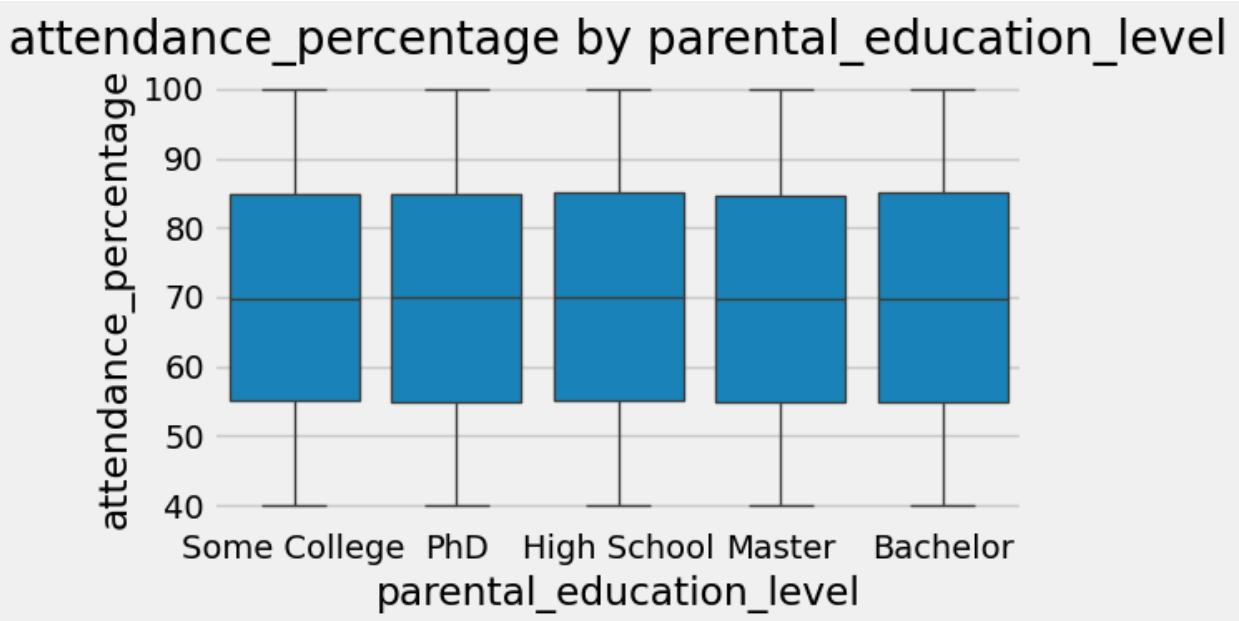
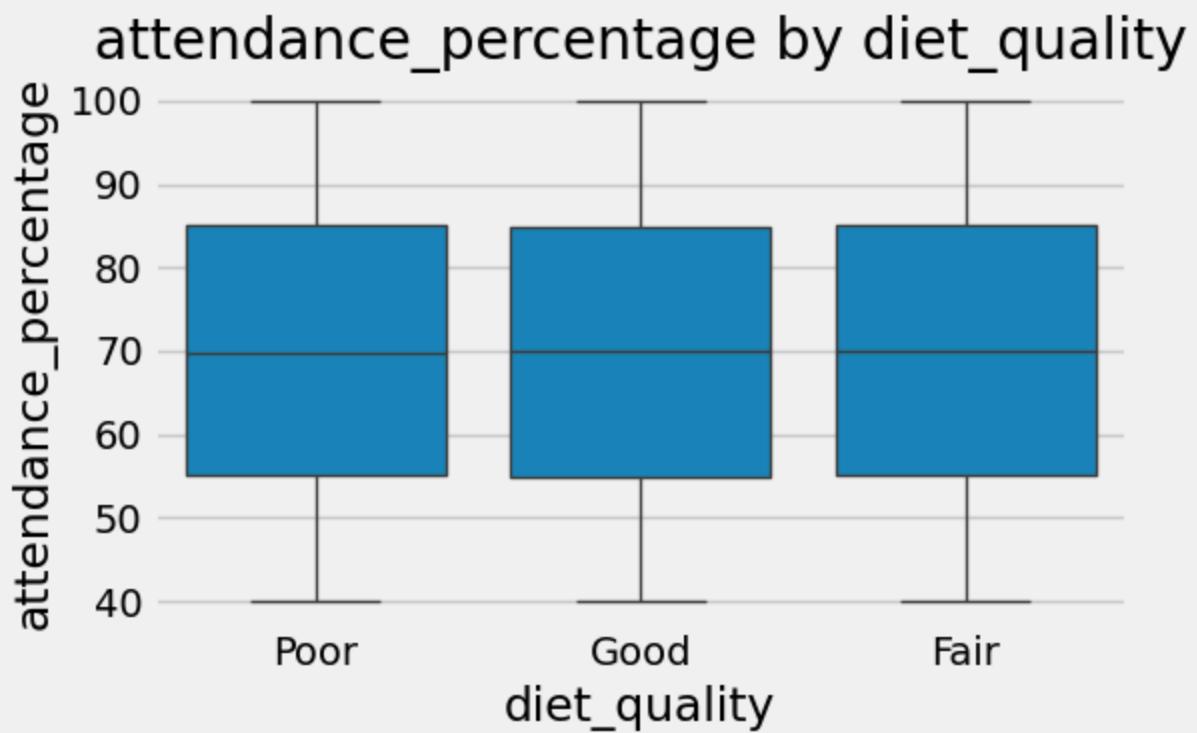
attendance_percentage vs exam_score (corr=0.00)

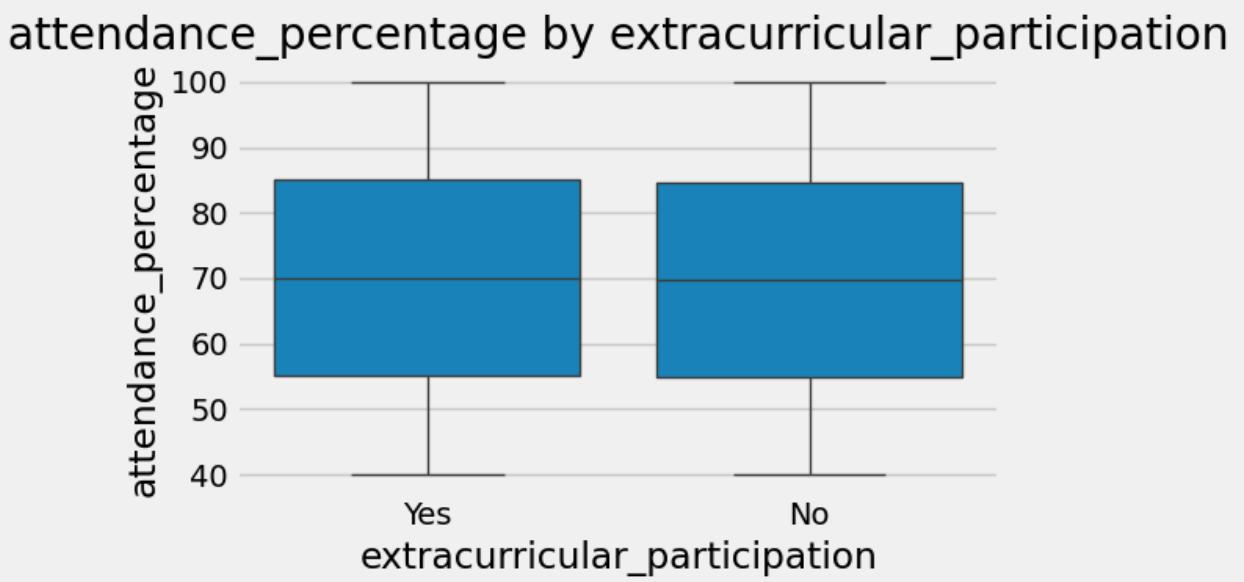
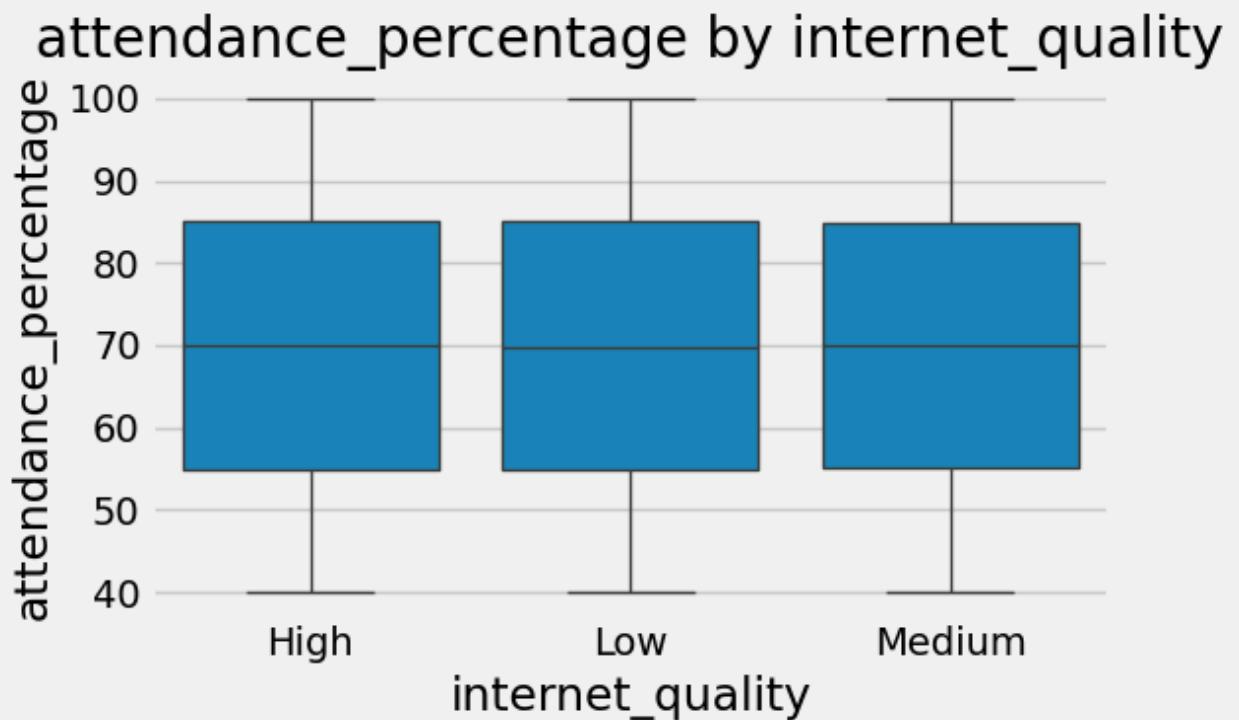


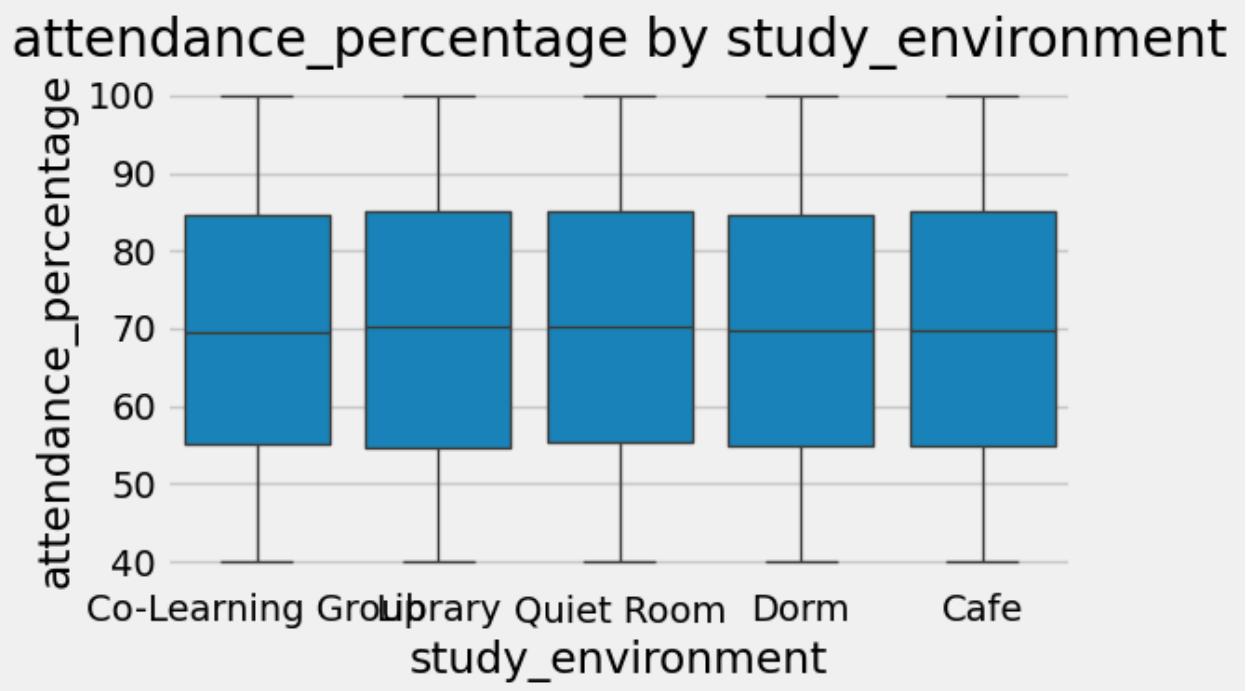
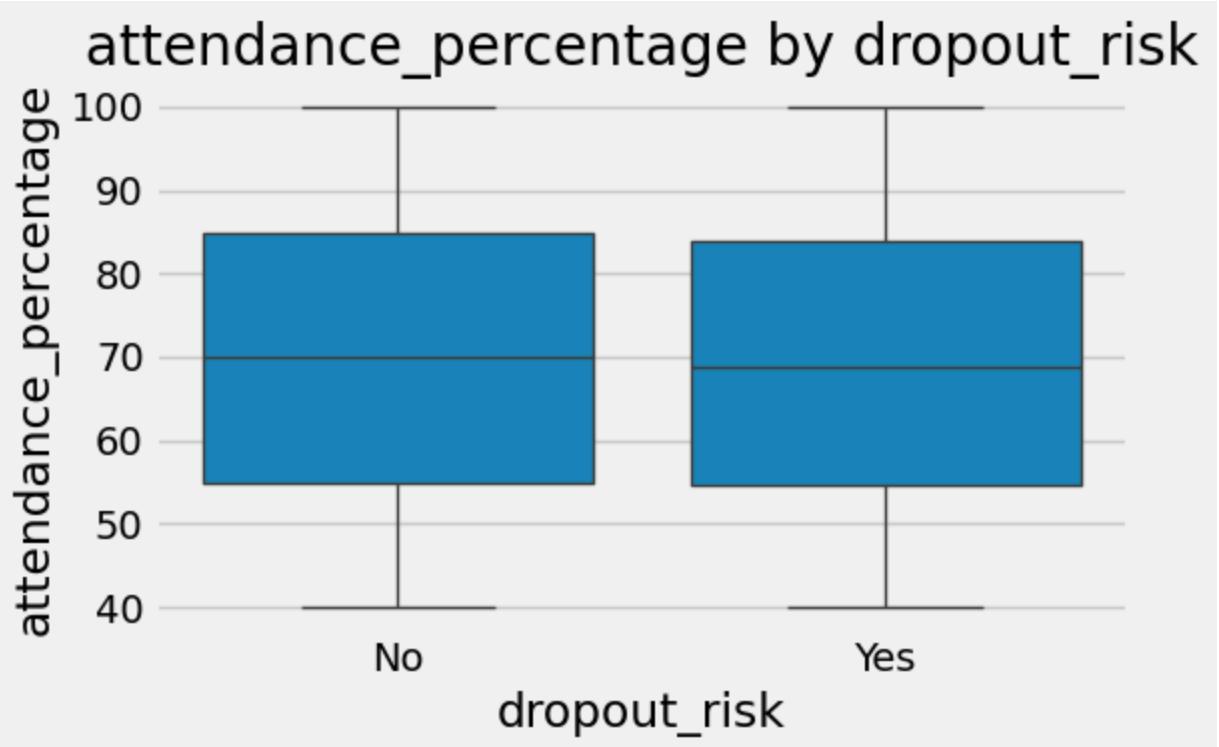
attendance_percentage by gender



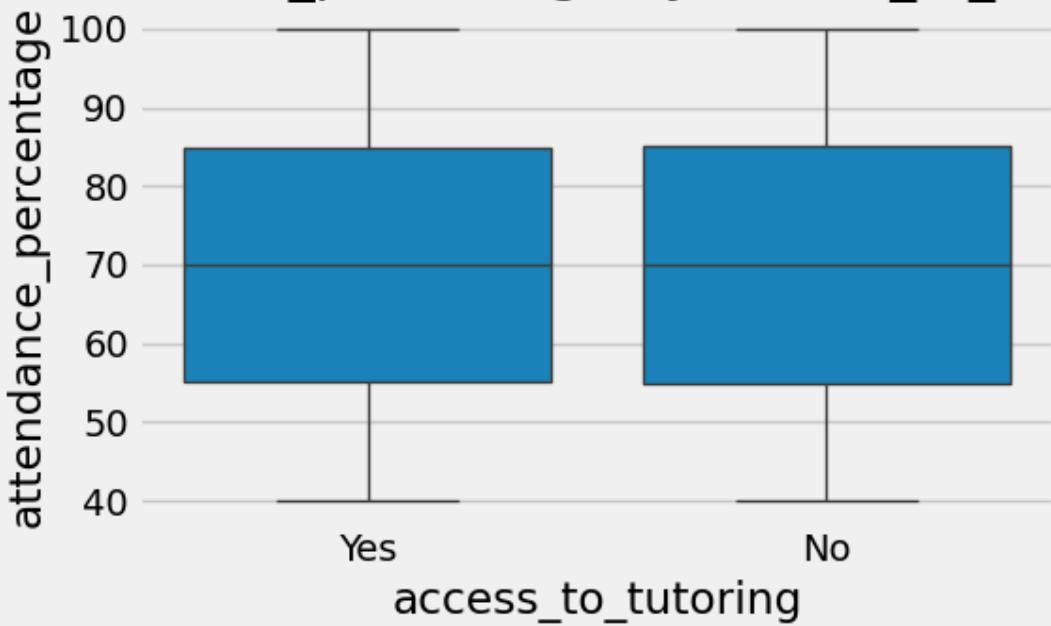




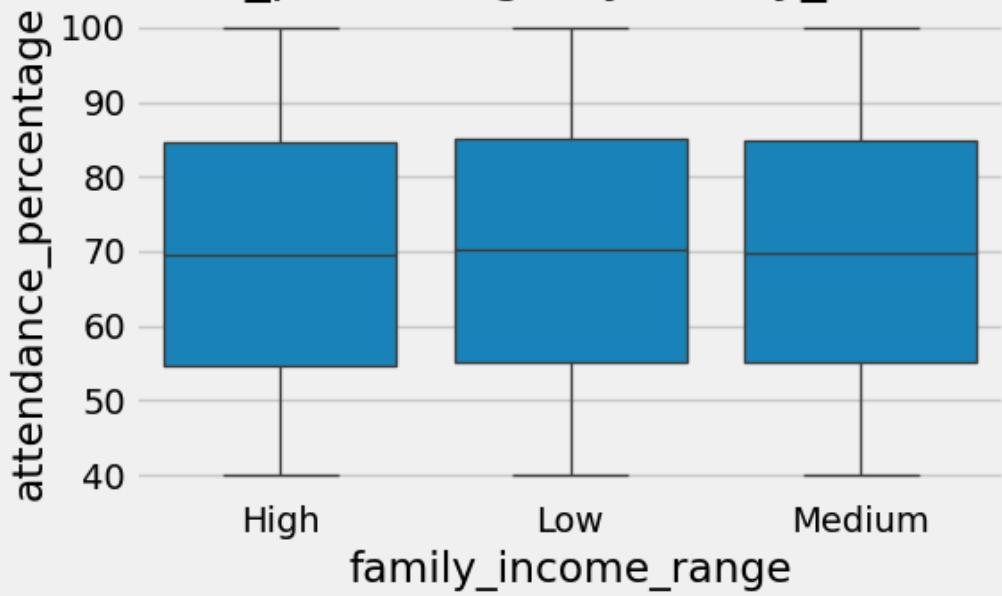


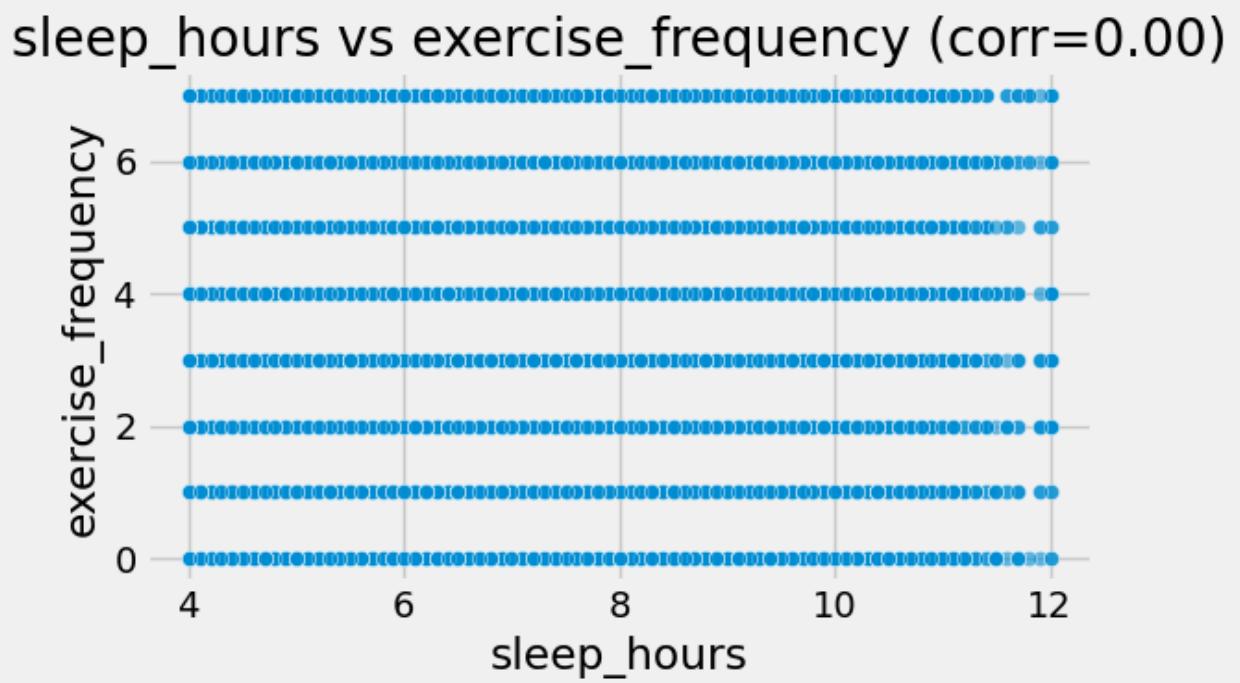
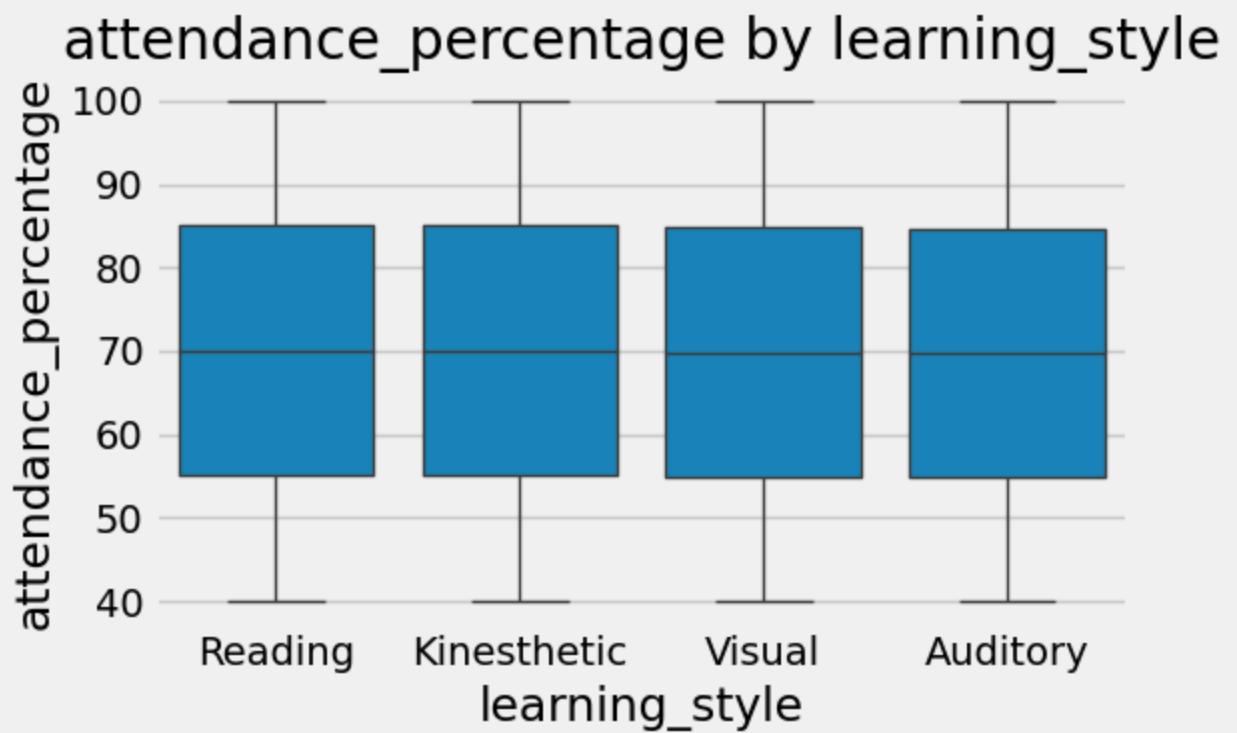


attendance_percentage by access_to_tutoring

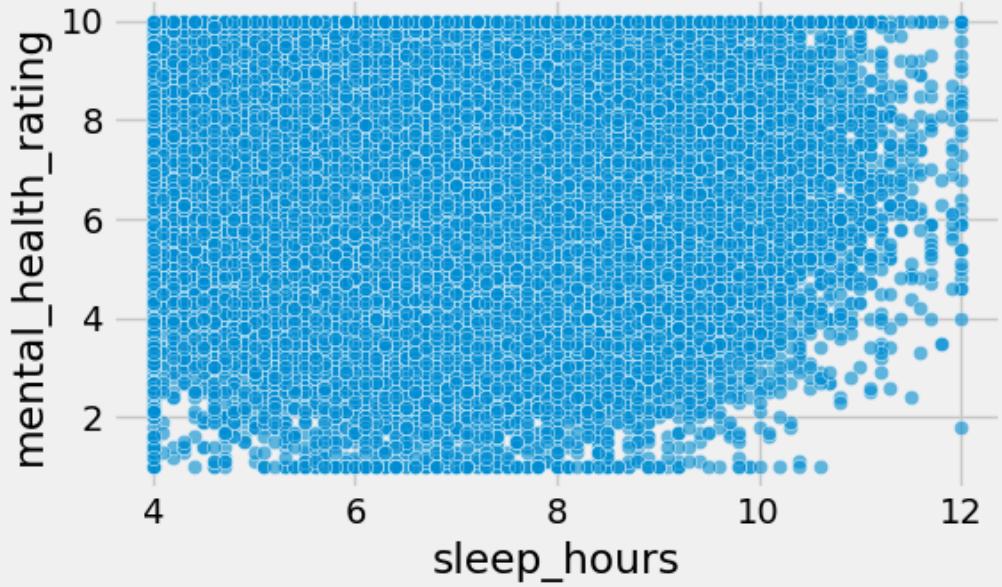


attendance_percentage by family_income_range

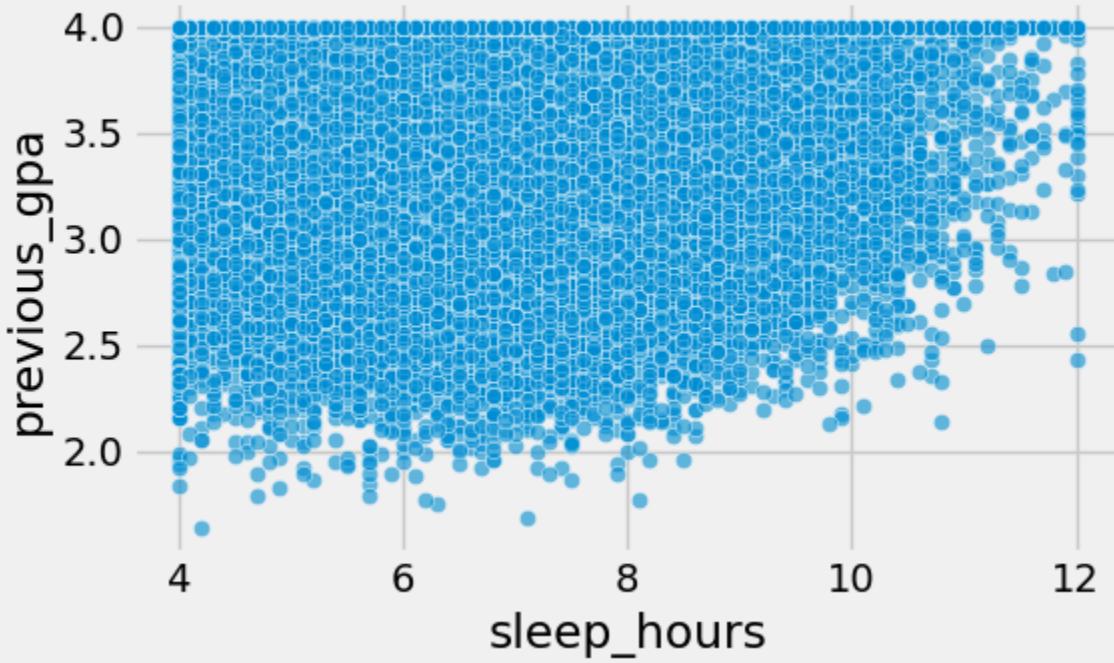


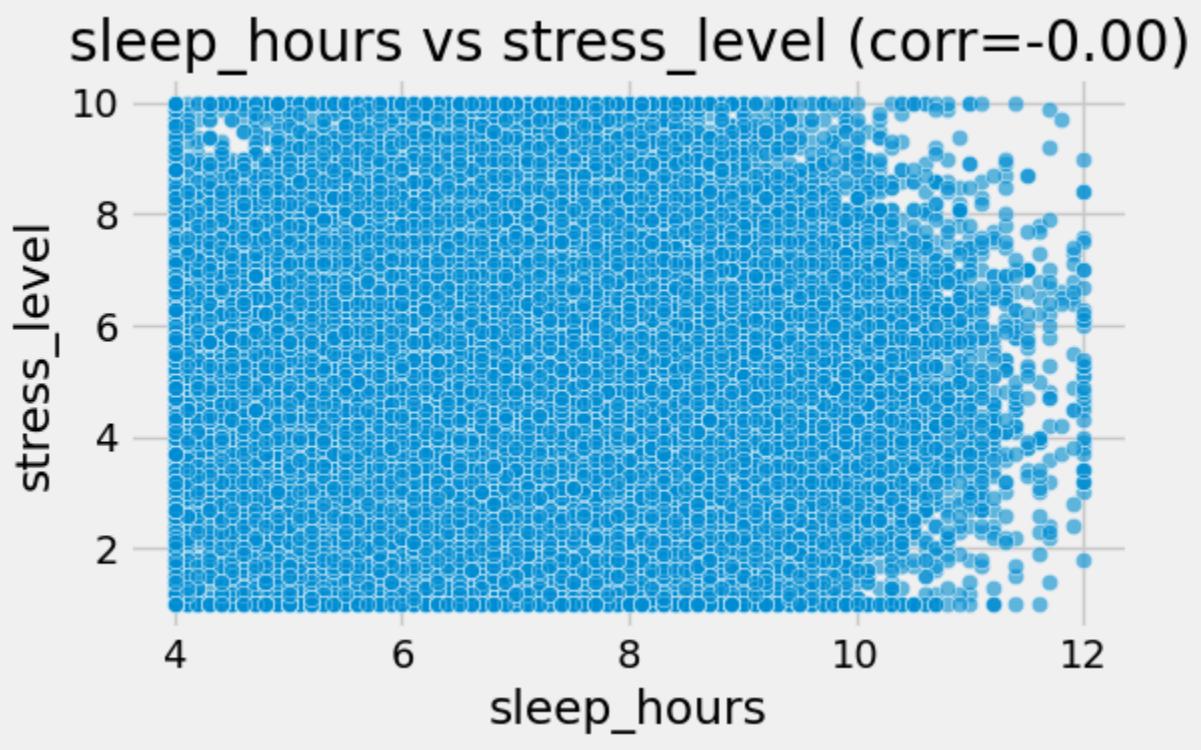
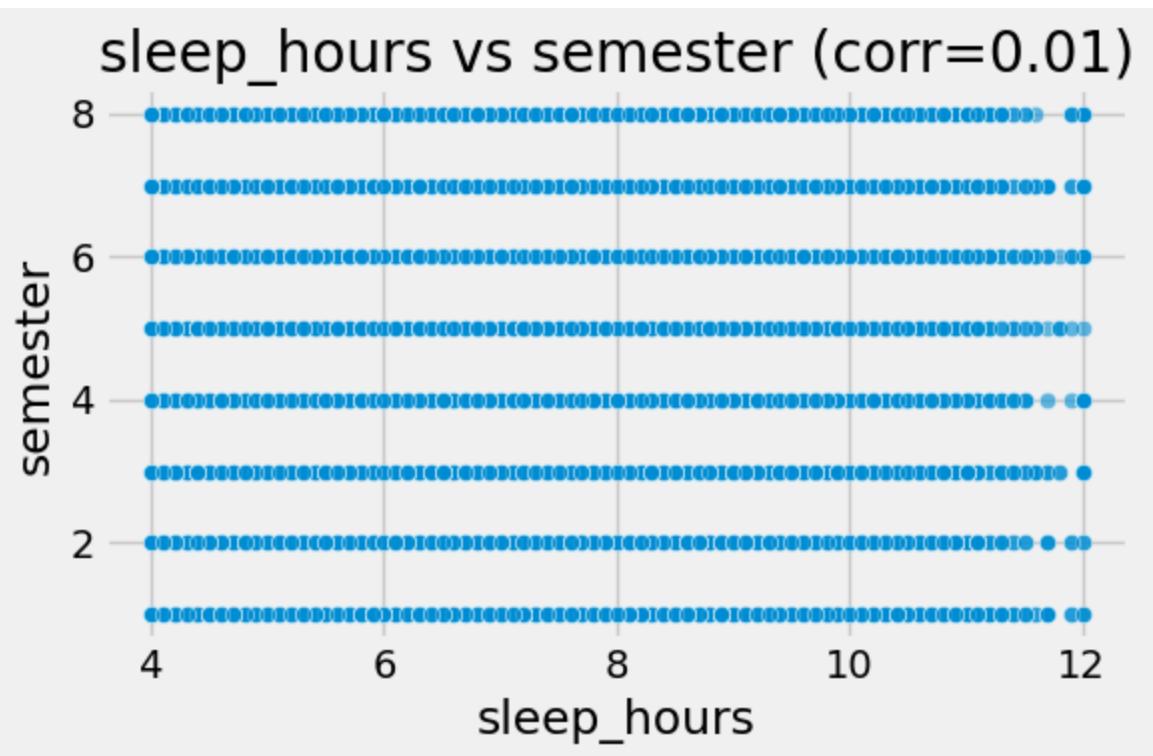


sleep_hours vs mental_health_rating (corr=0.01)

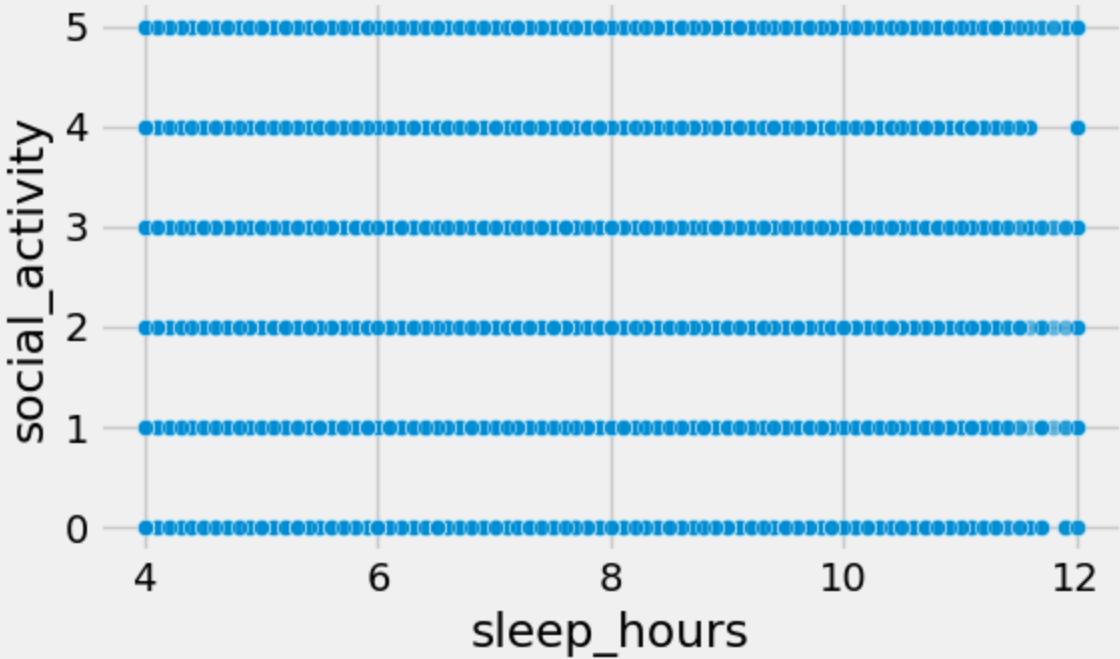


sleep_hours vs previous_gpa (corr=0.10)

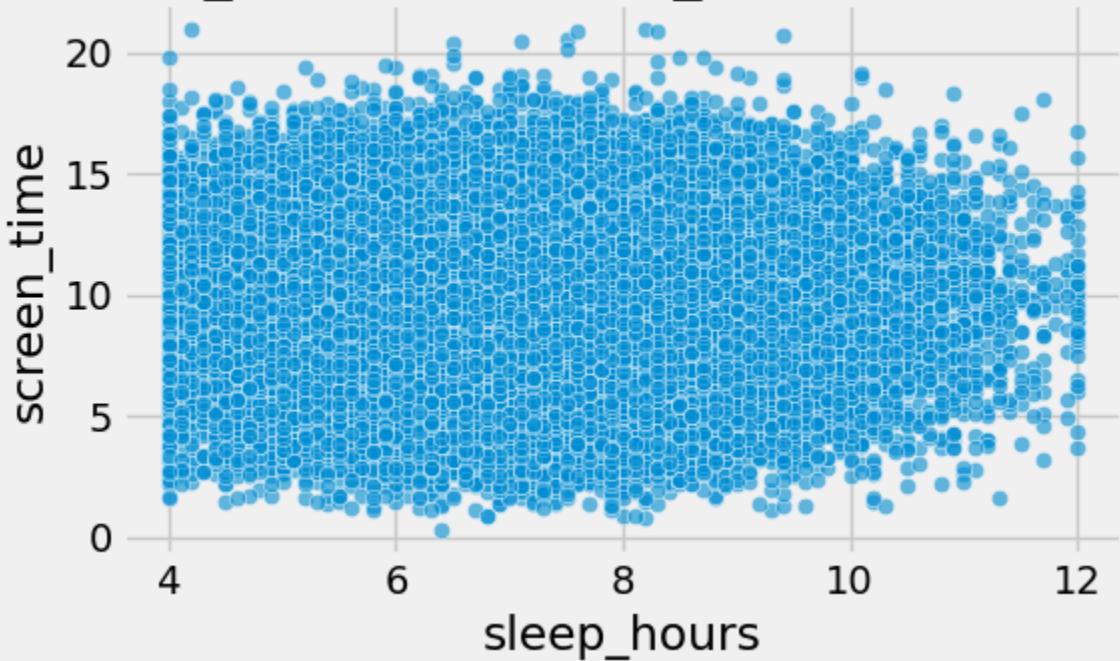




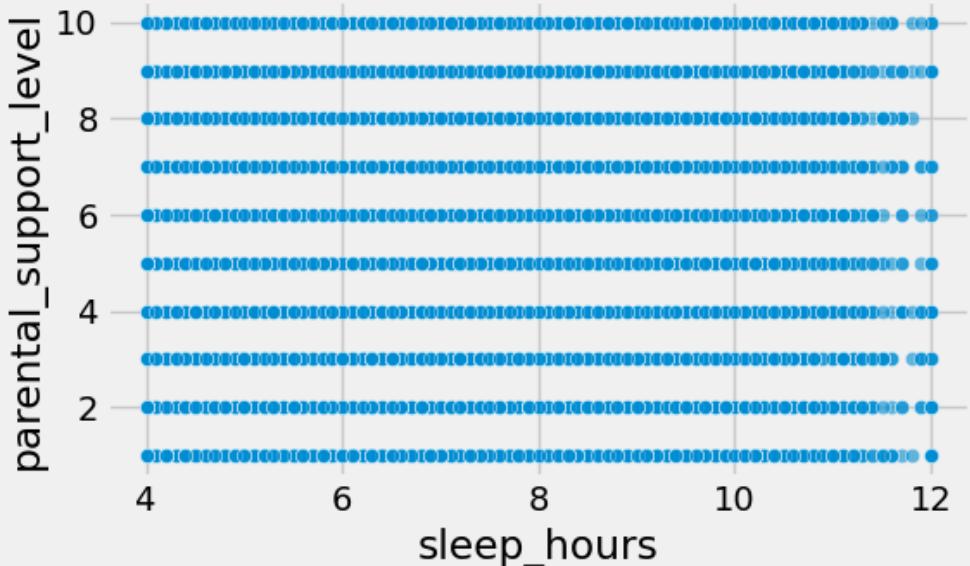
sleep_hours vs social_activity (corr=-0.00)



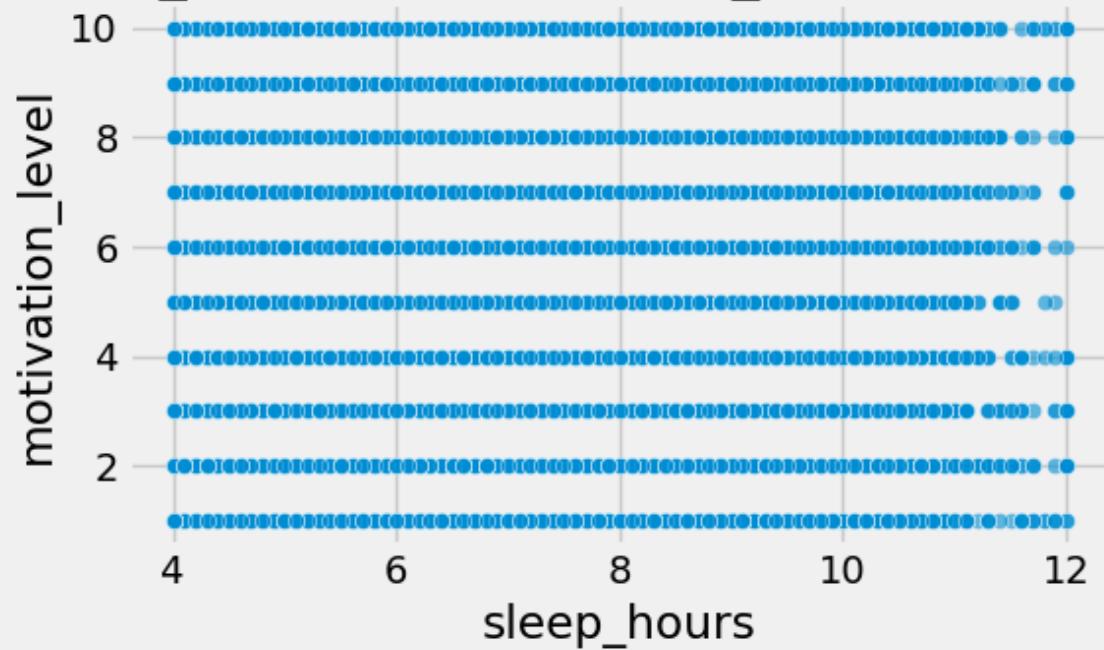
sleep_hours vs screen_time (corr=-0.00)



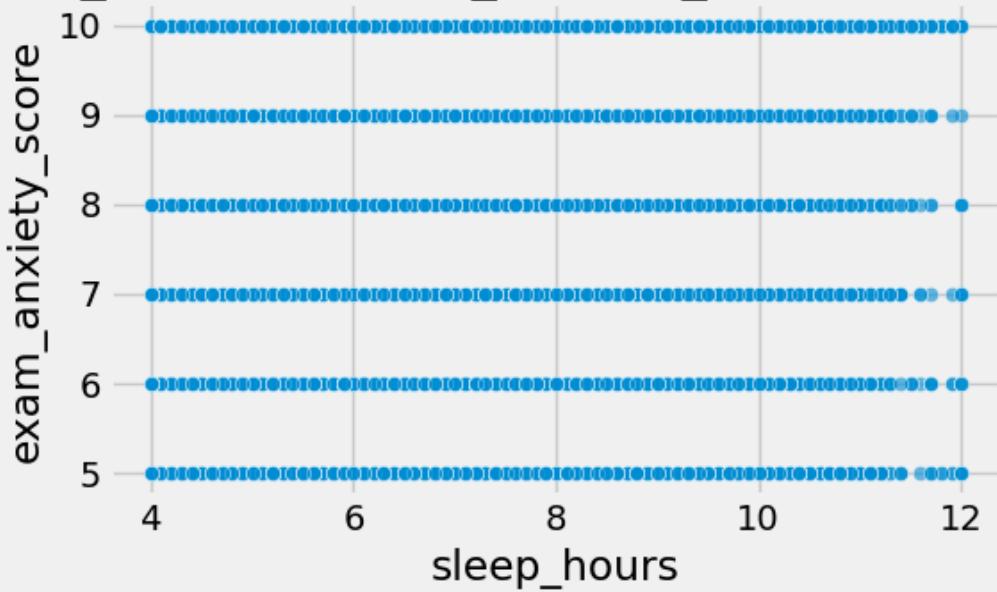
sleep_hours vs parental_support_level (corr=0.00)



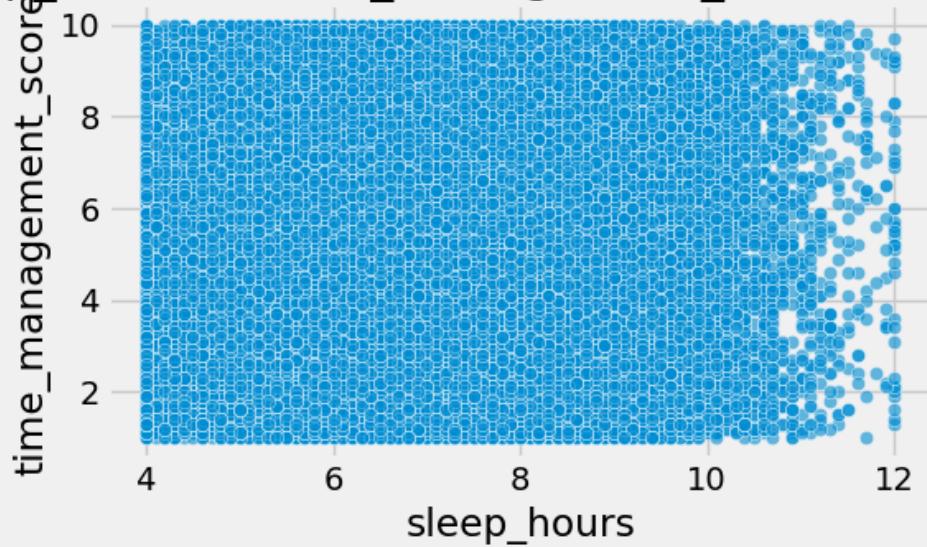
sleep_hours vs motivation_level (corr=0.00)

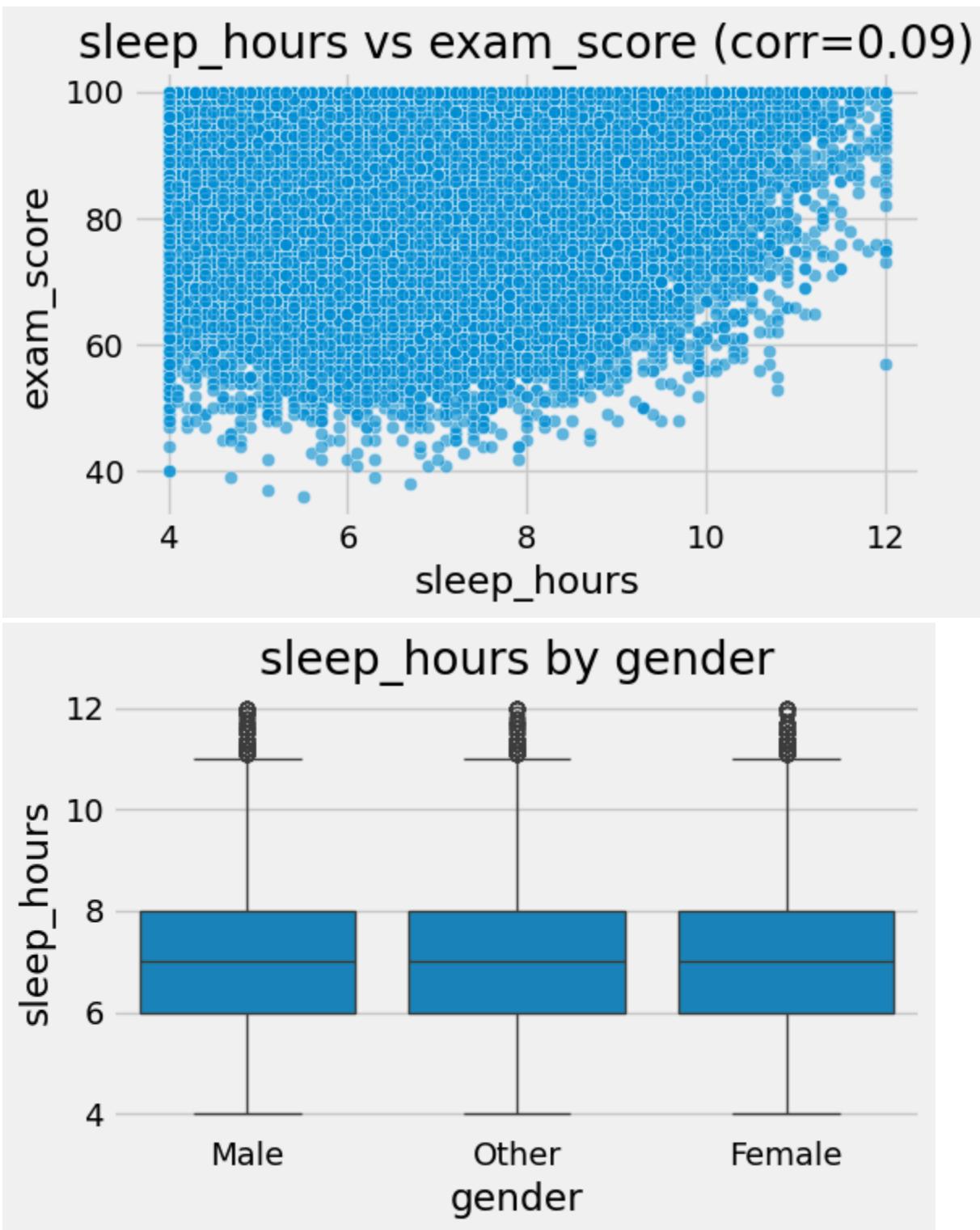


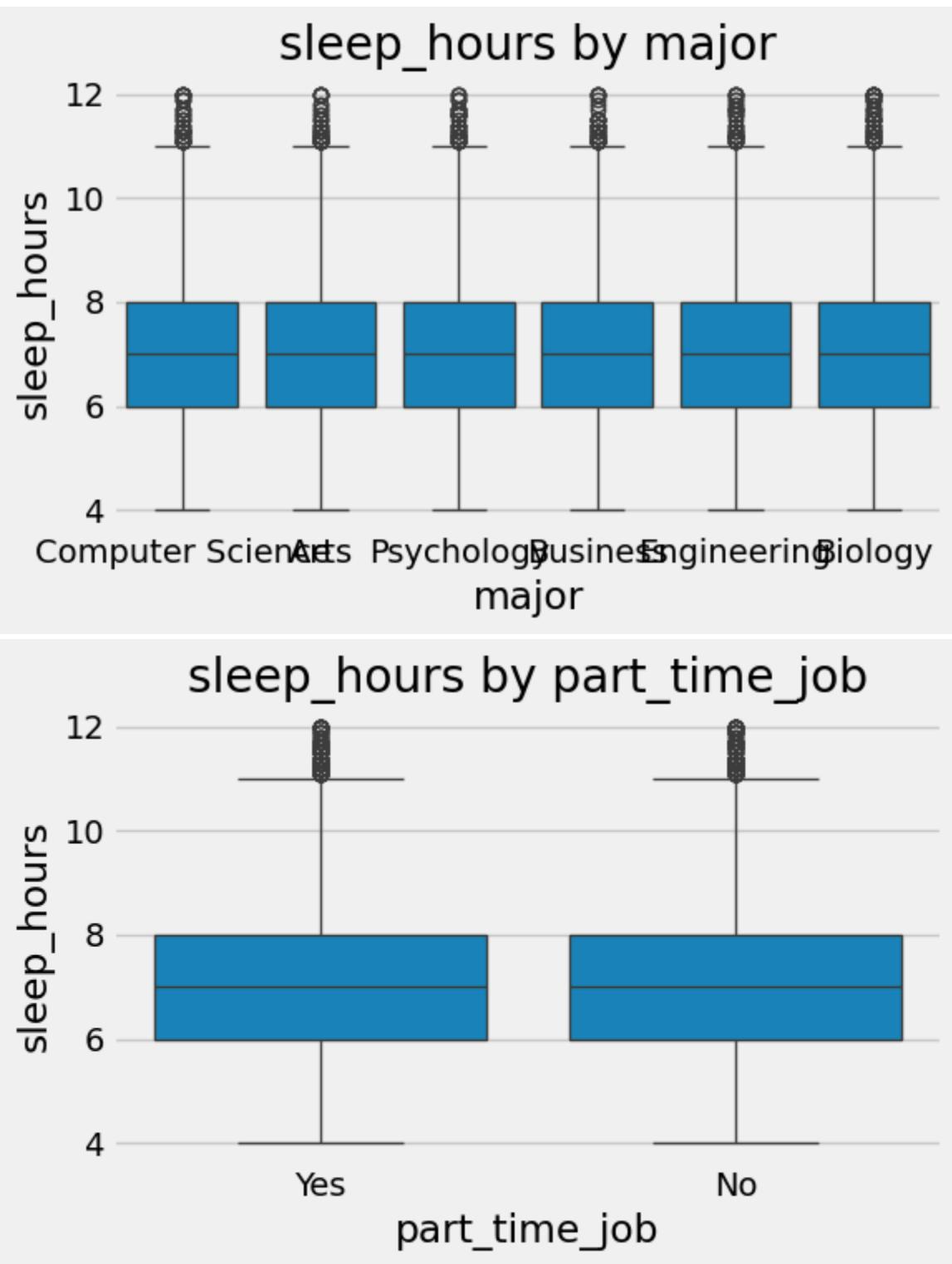
sleep_hours vs exam_anxiety_score (corr=-0.00)

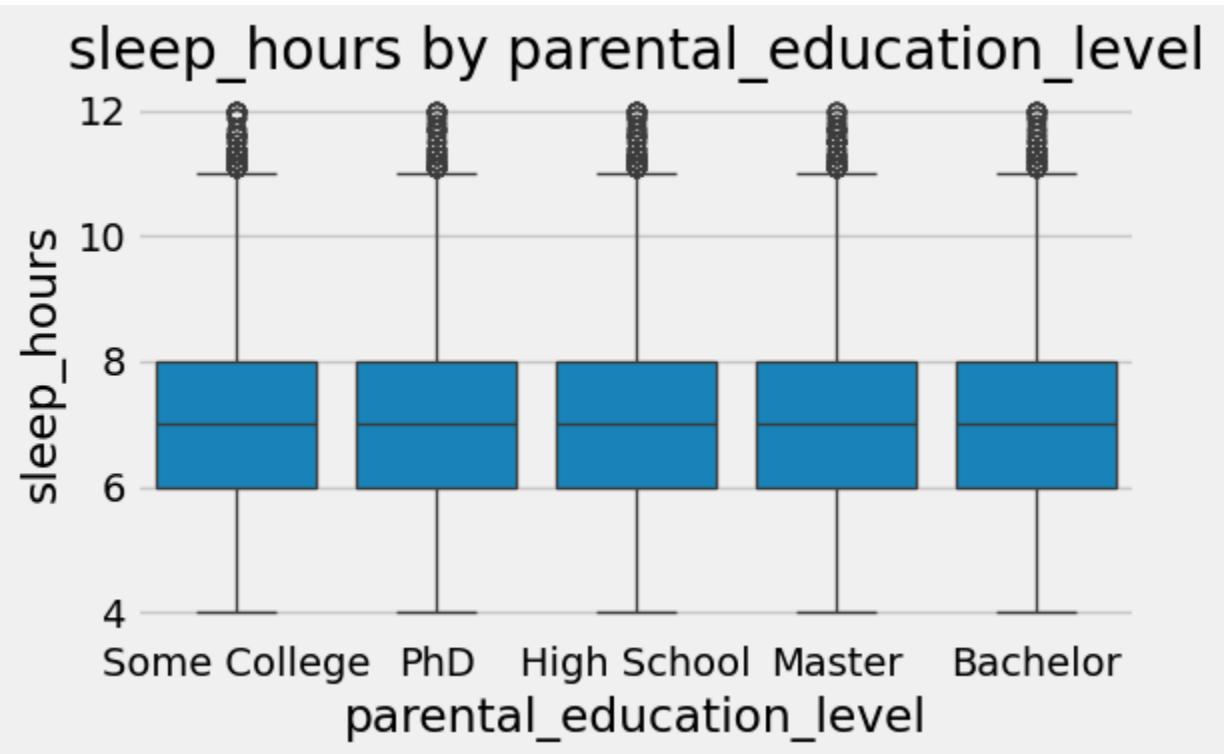
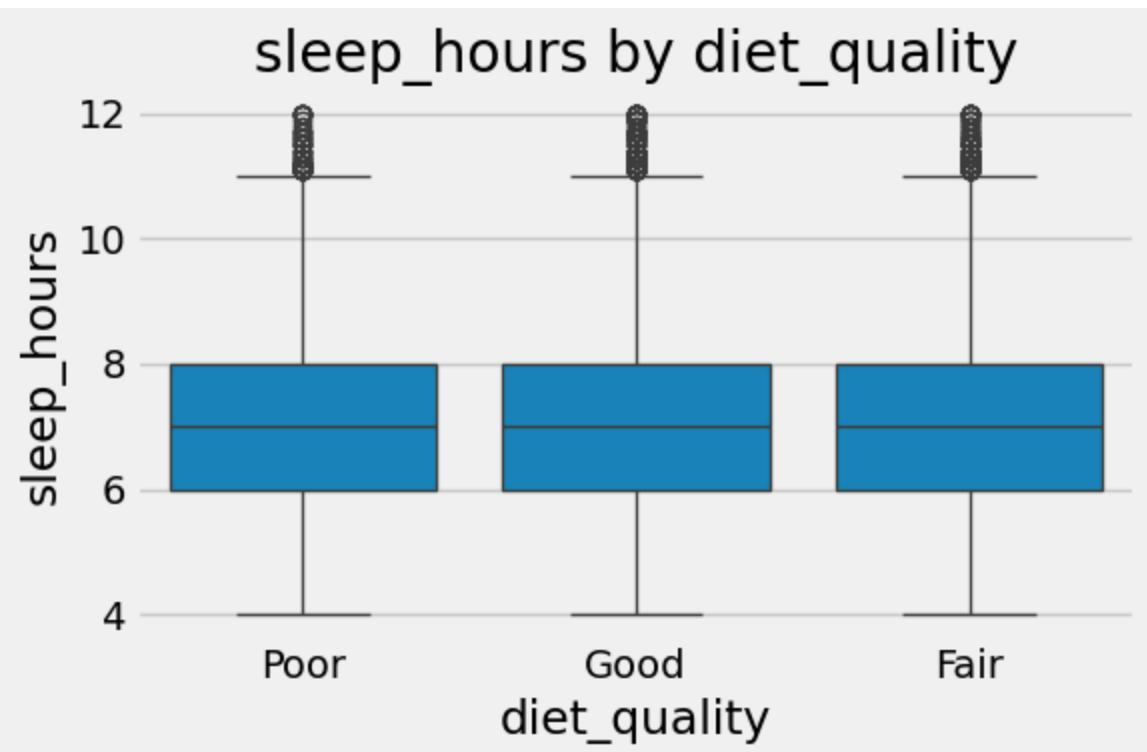


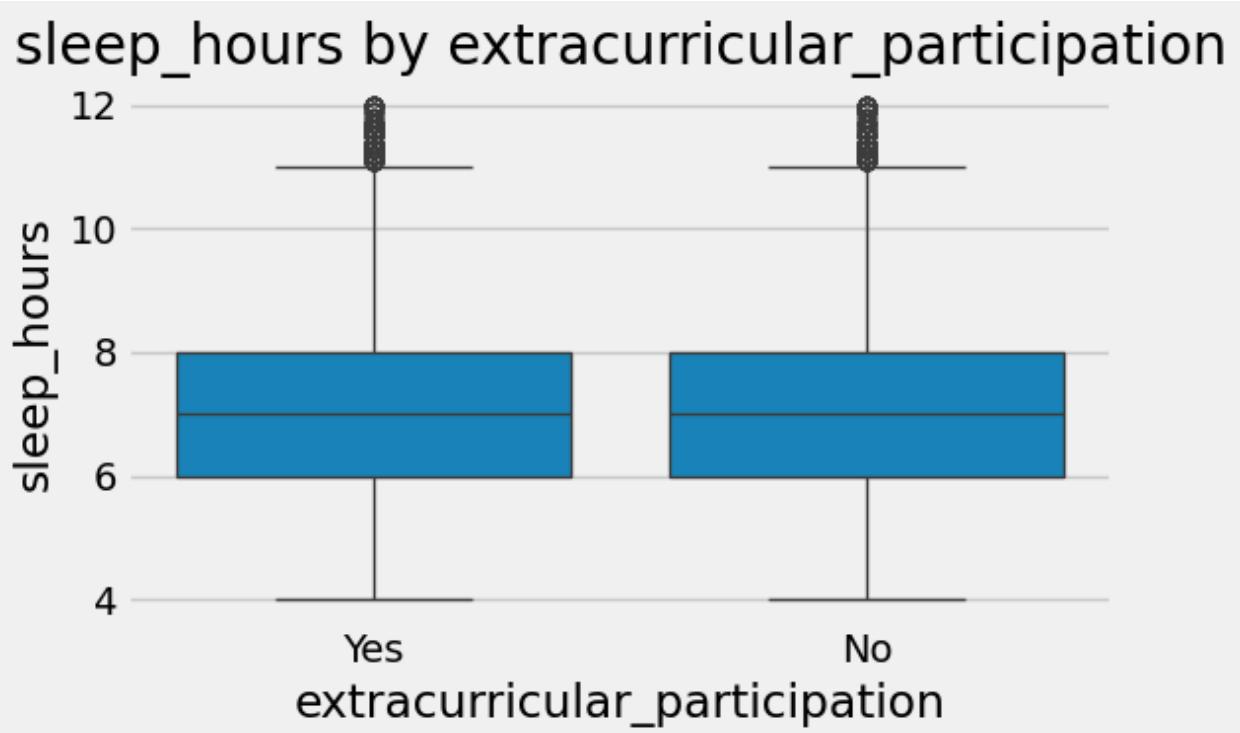
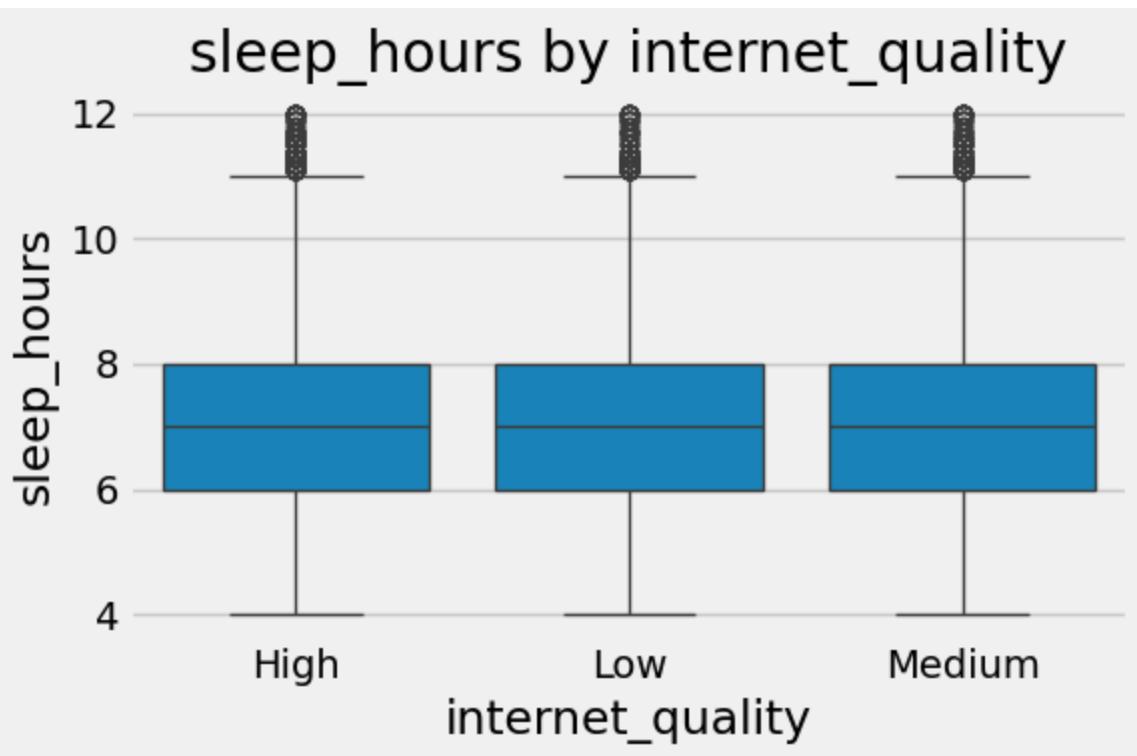
sleep_hours vs time_management_score (corr=0.00)

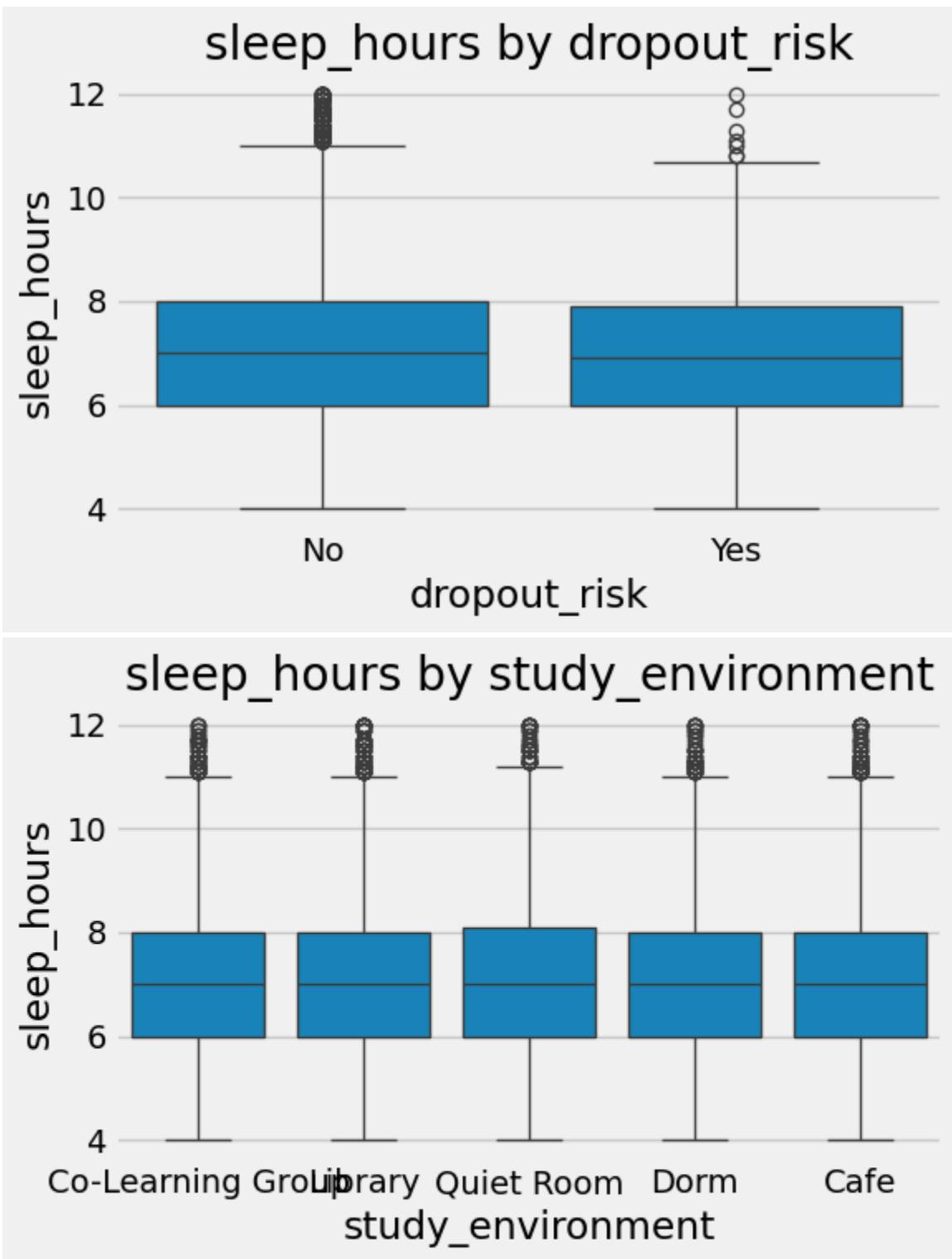


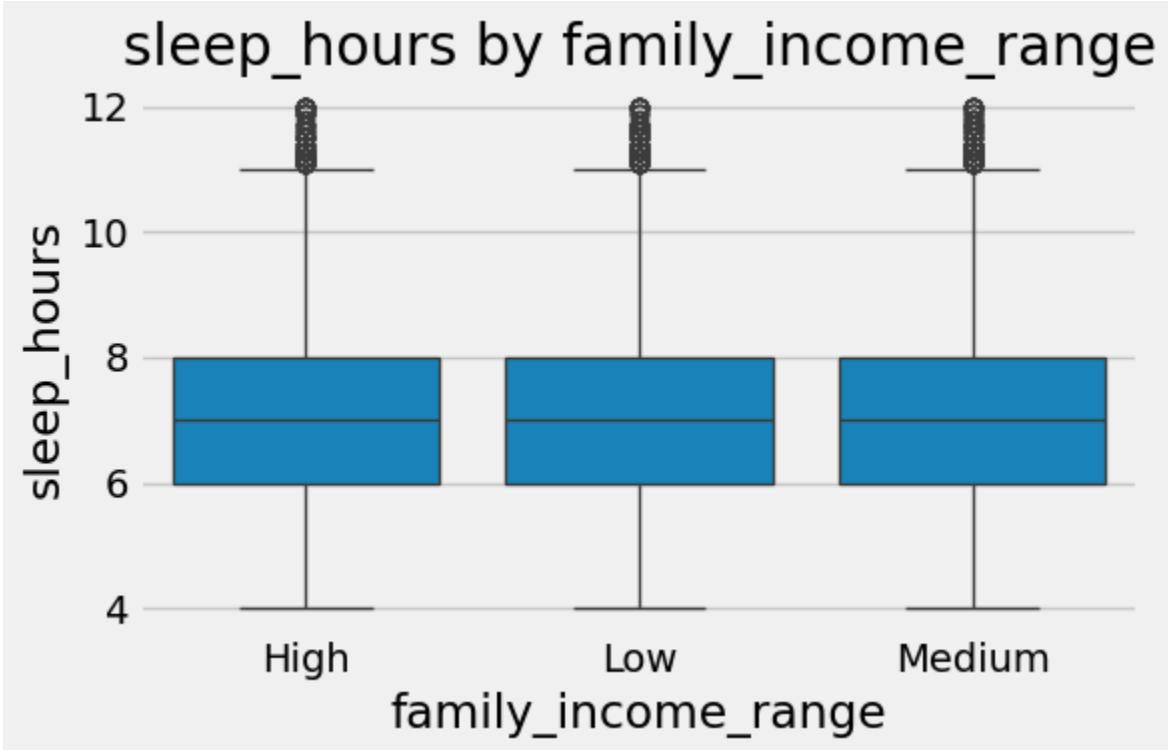
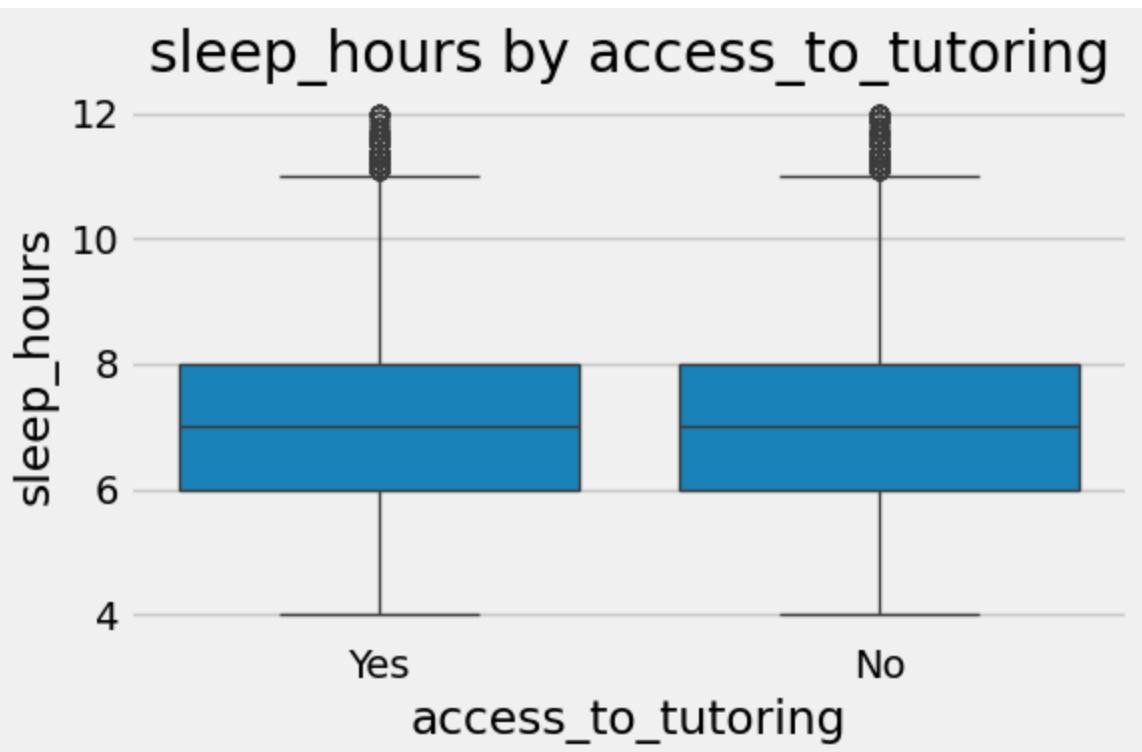


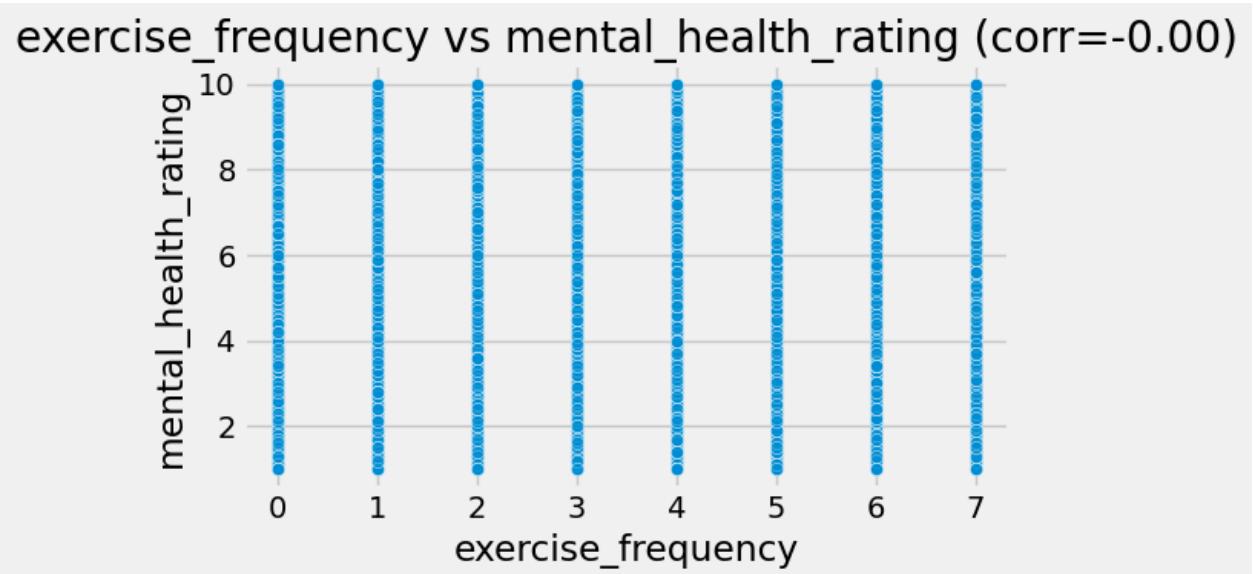
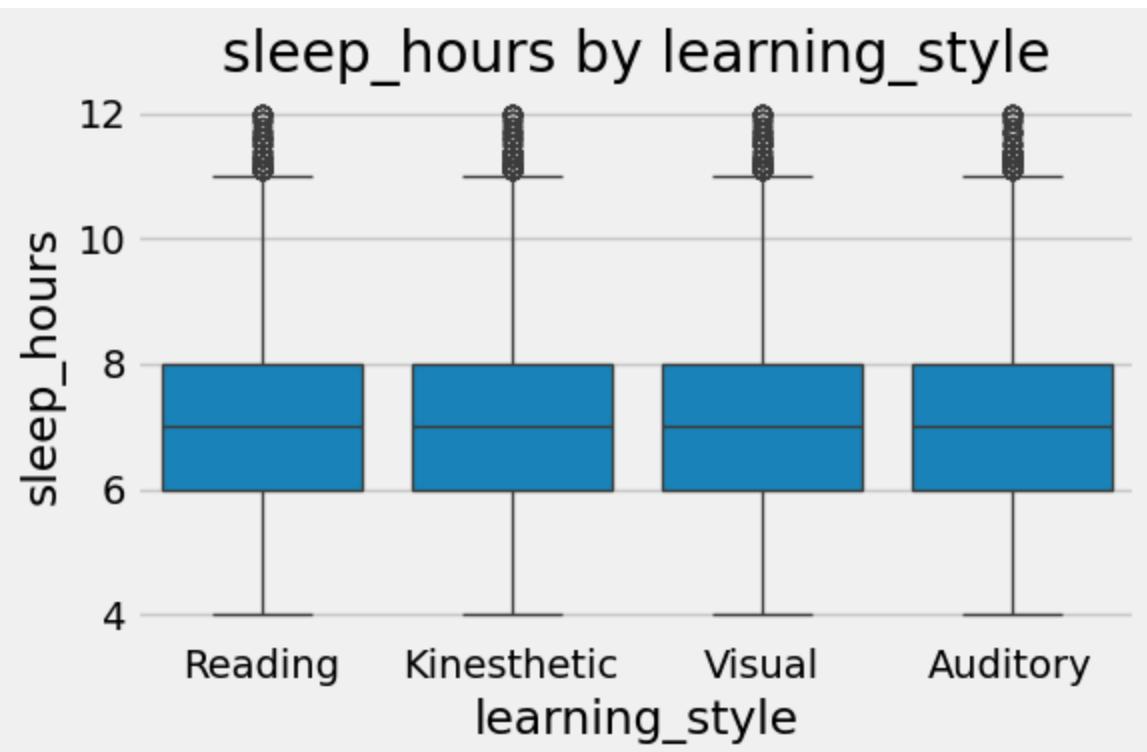




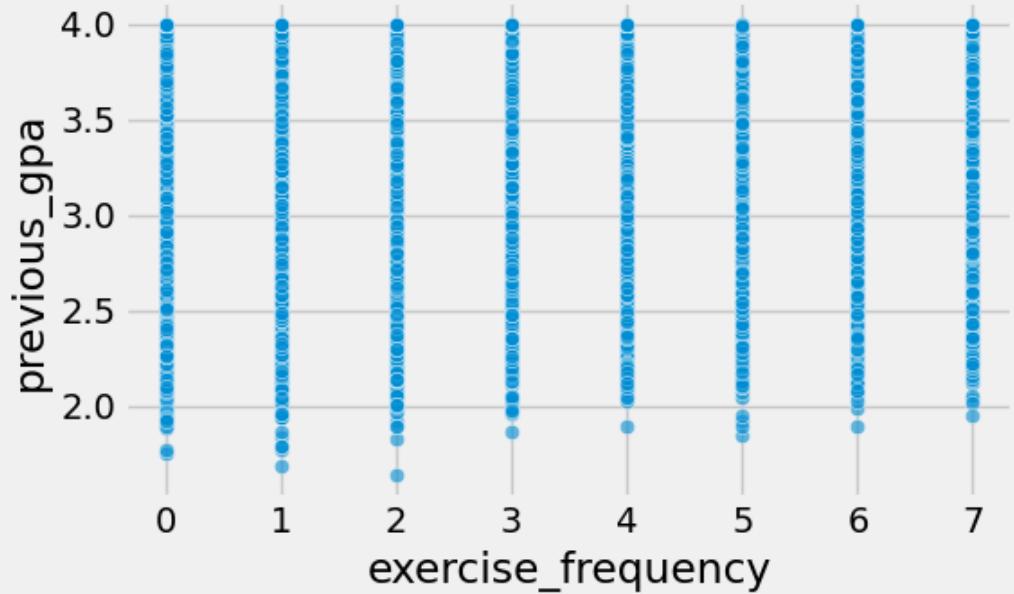




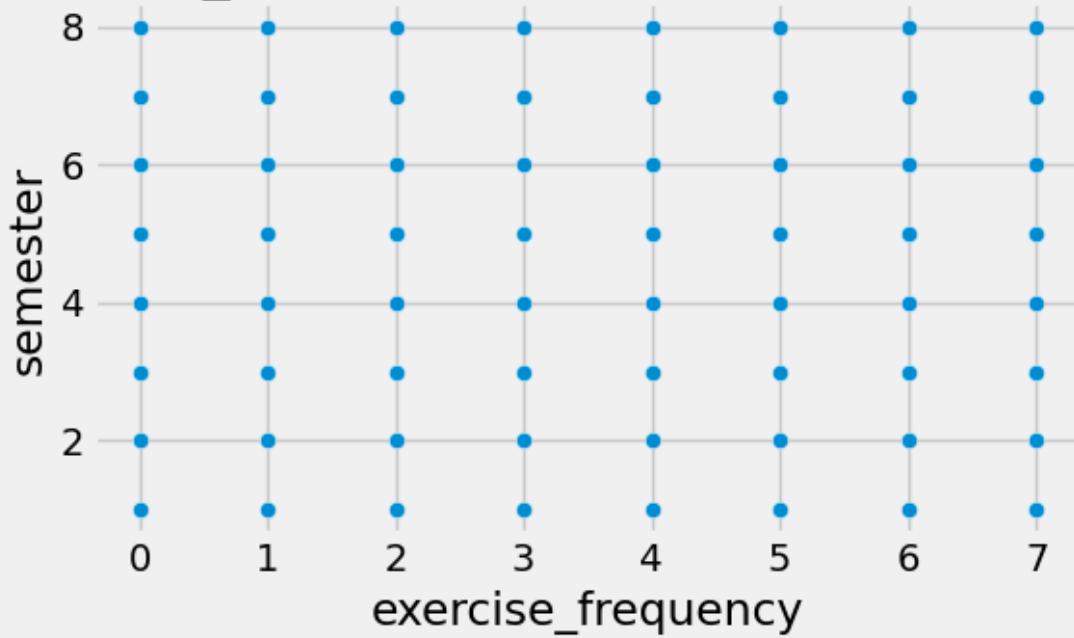




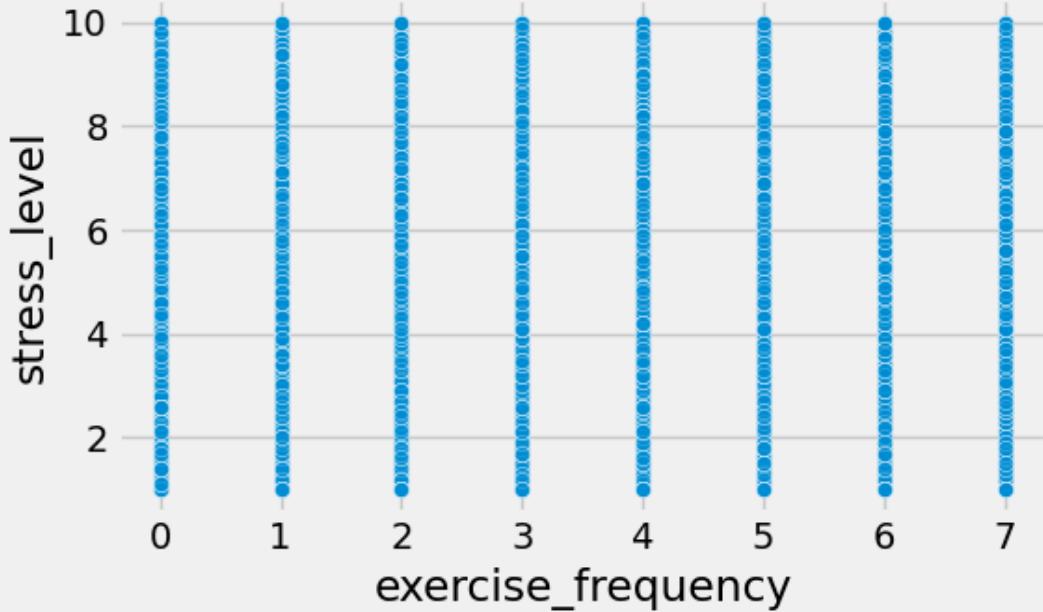
exercise_frequency vs previous_gpa (corr=0.09)



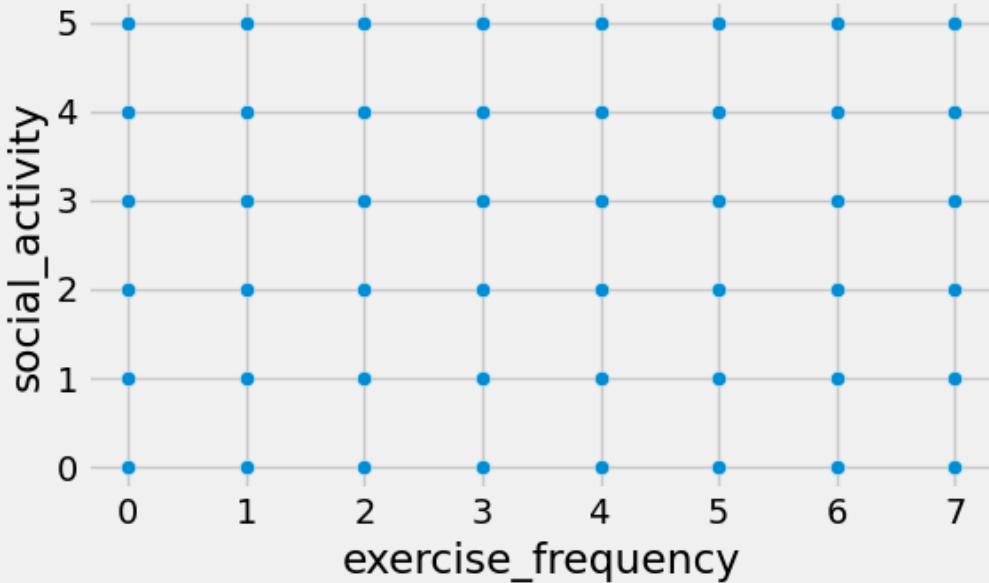
exercise_frequency vs semester (corr=-0.00)



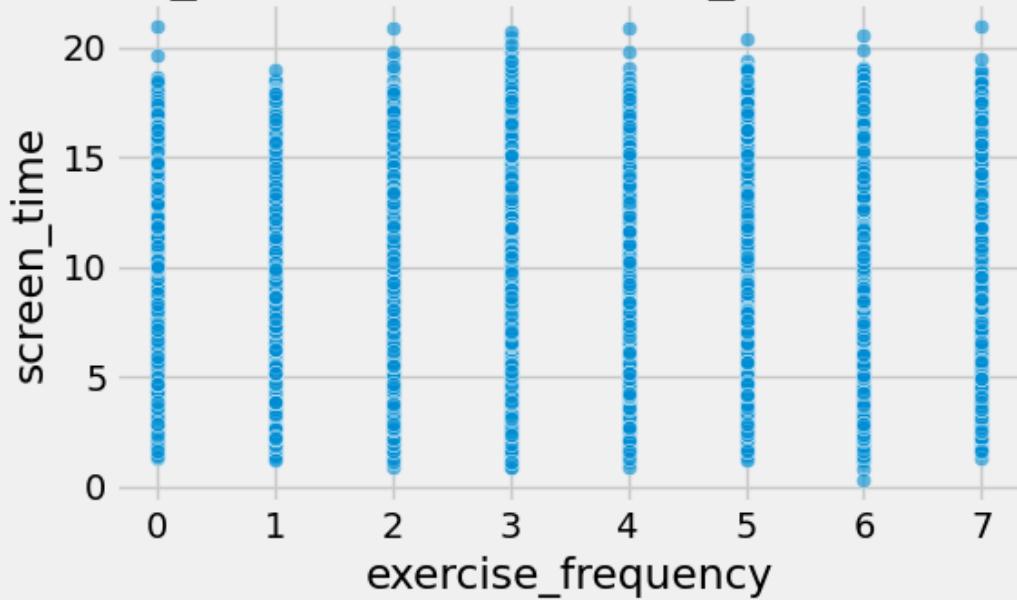
exercise_frequency vs stress_level (corr=0.01)



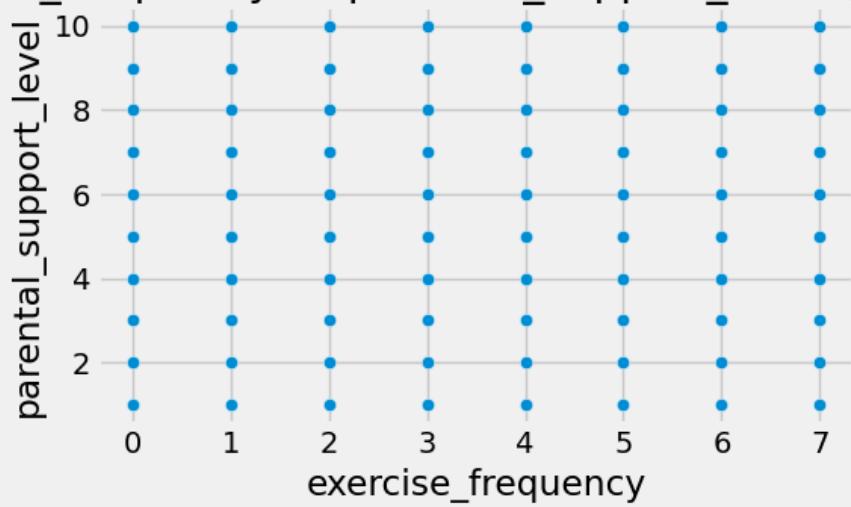
exercise_frequency vs social_activity (corr=0.00)



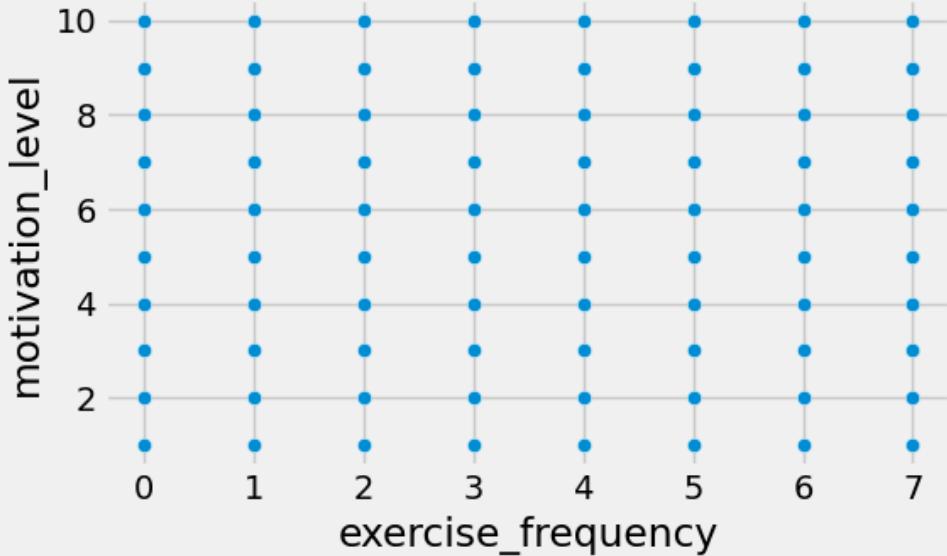
exercise_frequency vs screen_time (corr=-0.00)



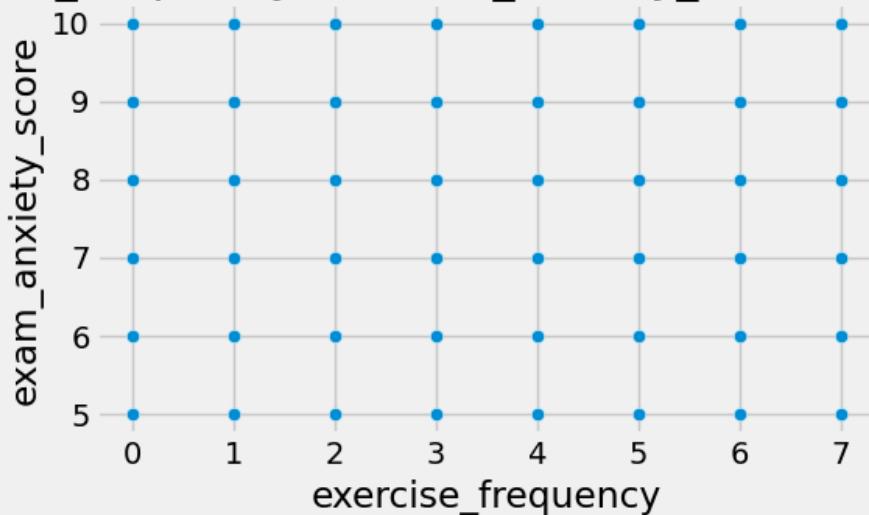
exercise_frequency vs parental_support_level (corr=0.00)



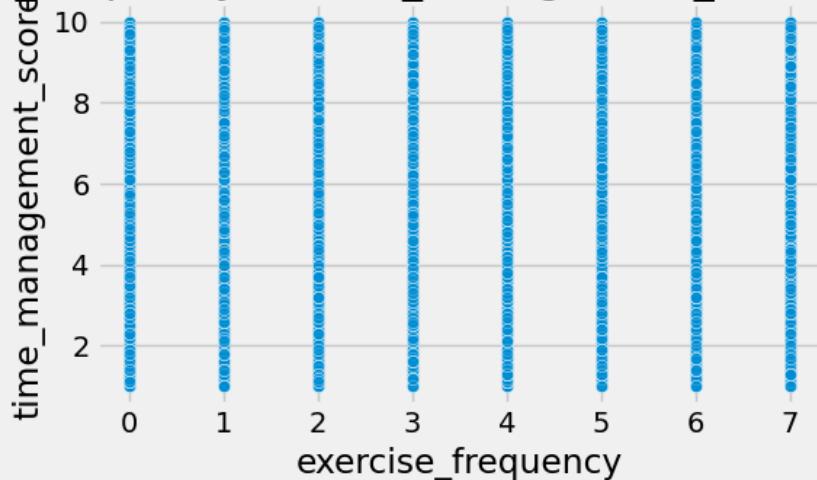
exercise_frequency vs motivation_level (corr=0.01)



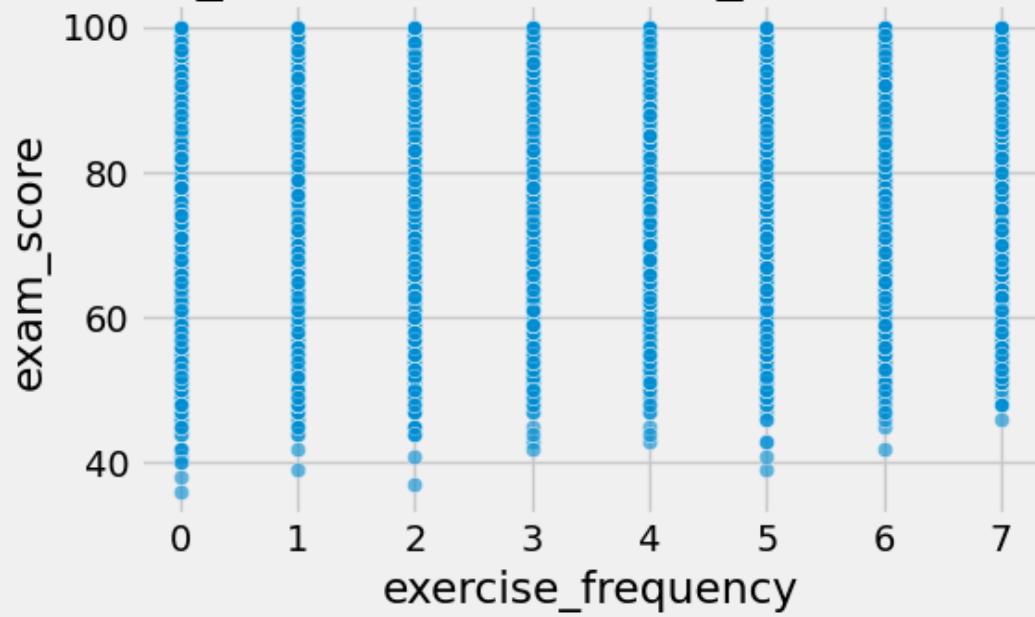
exercise_frequency vs exam_anxiety_score (corr=-0.01)



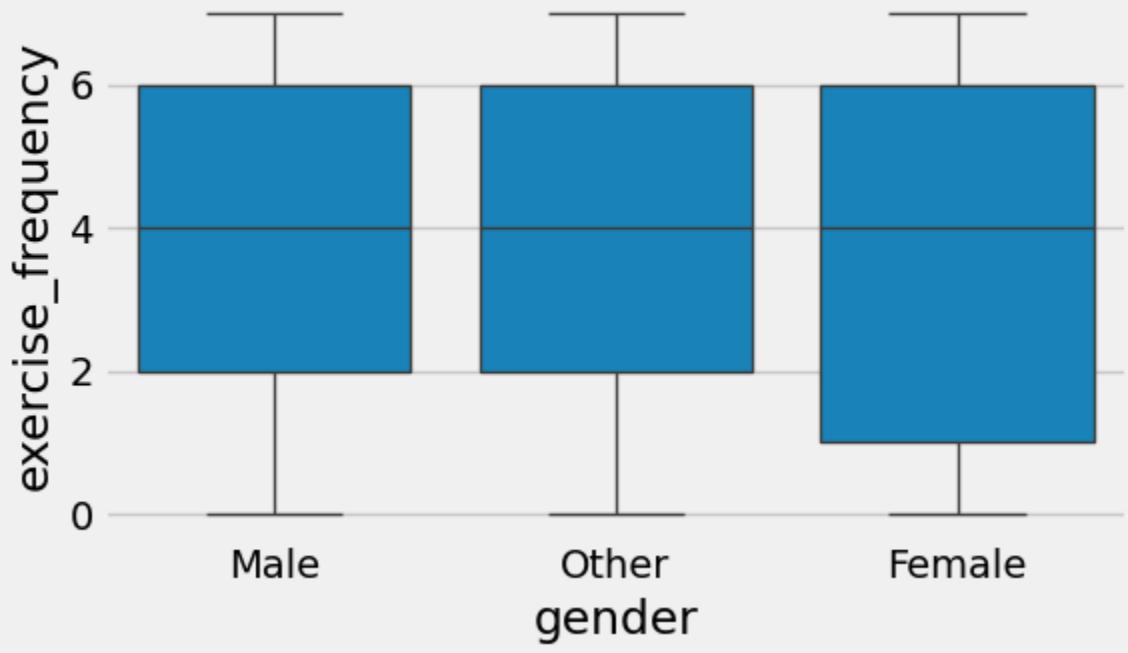
exercise_frequency vs time_management_score (corr=0.00)

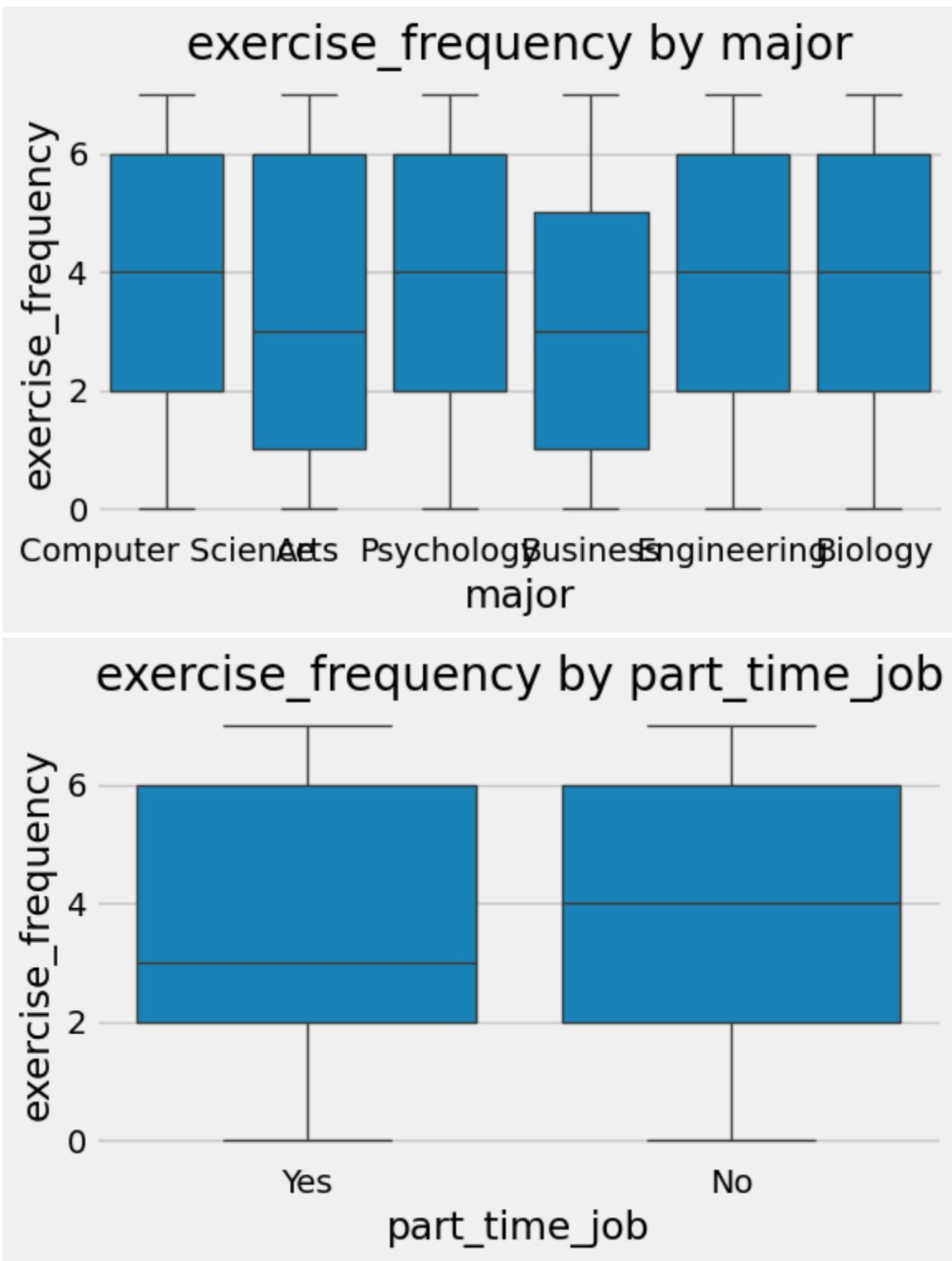


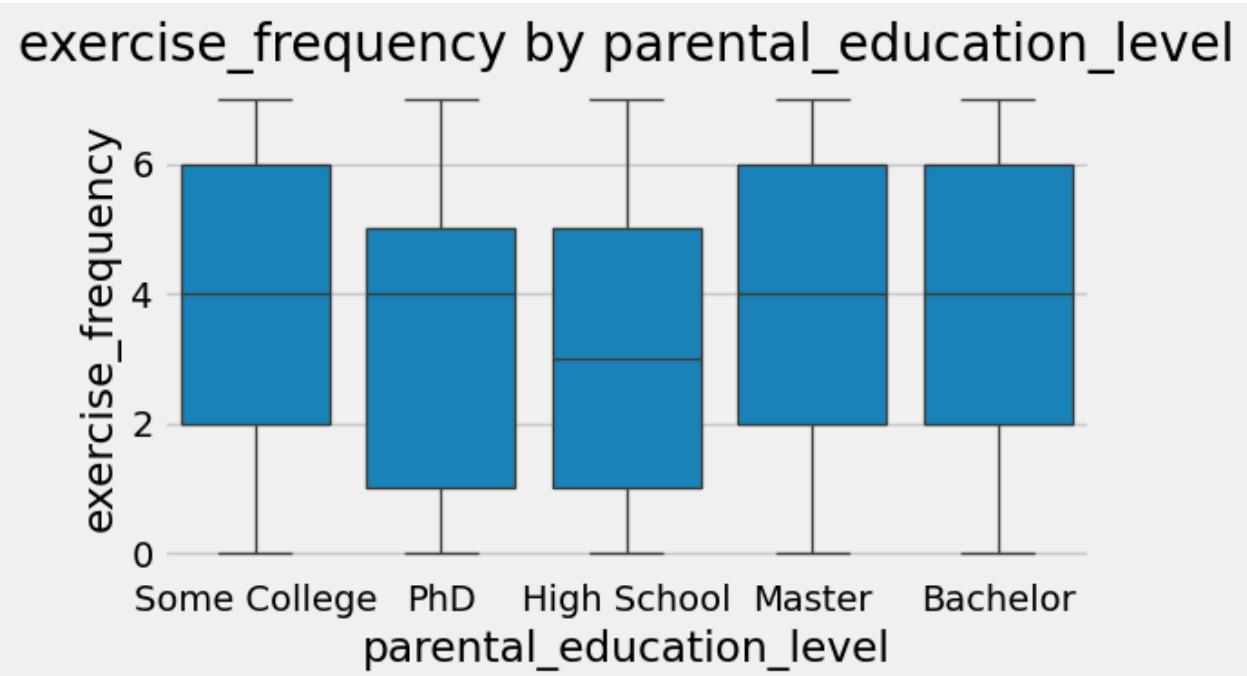
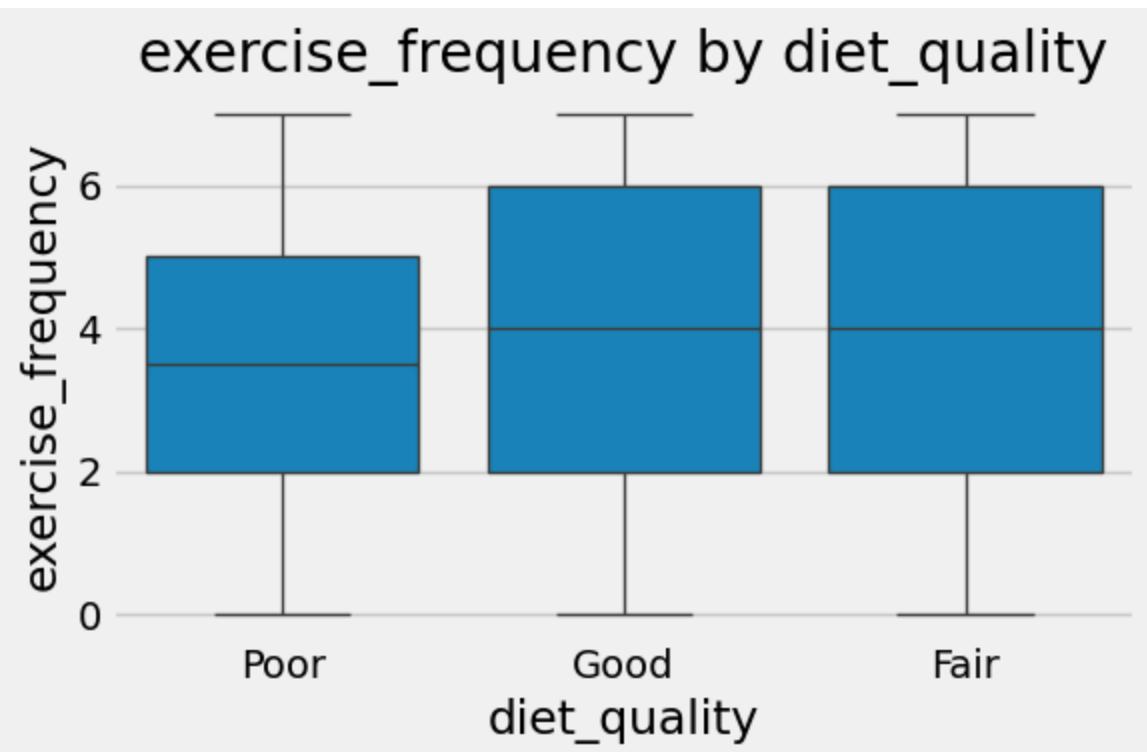
exercise_frequency vs exam_score (corr=0.09)



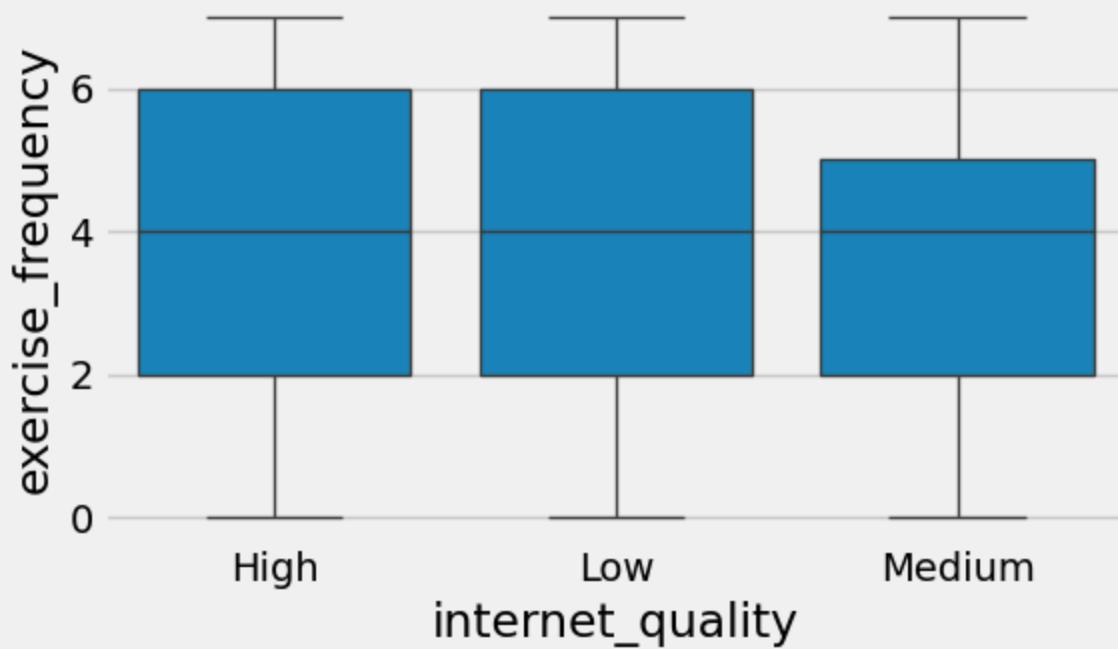
exercise_frequency by gender



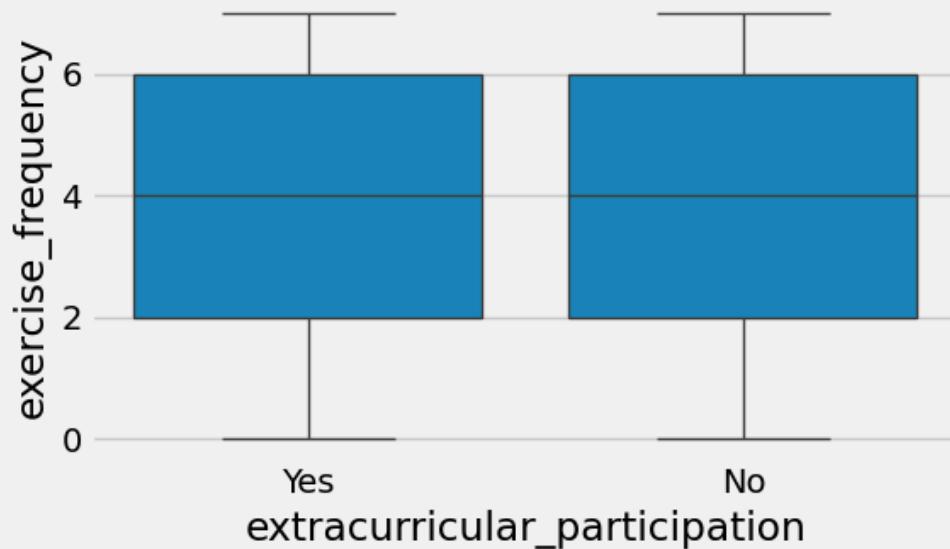


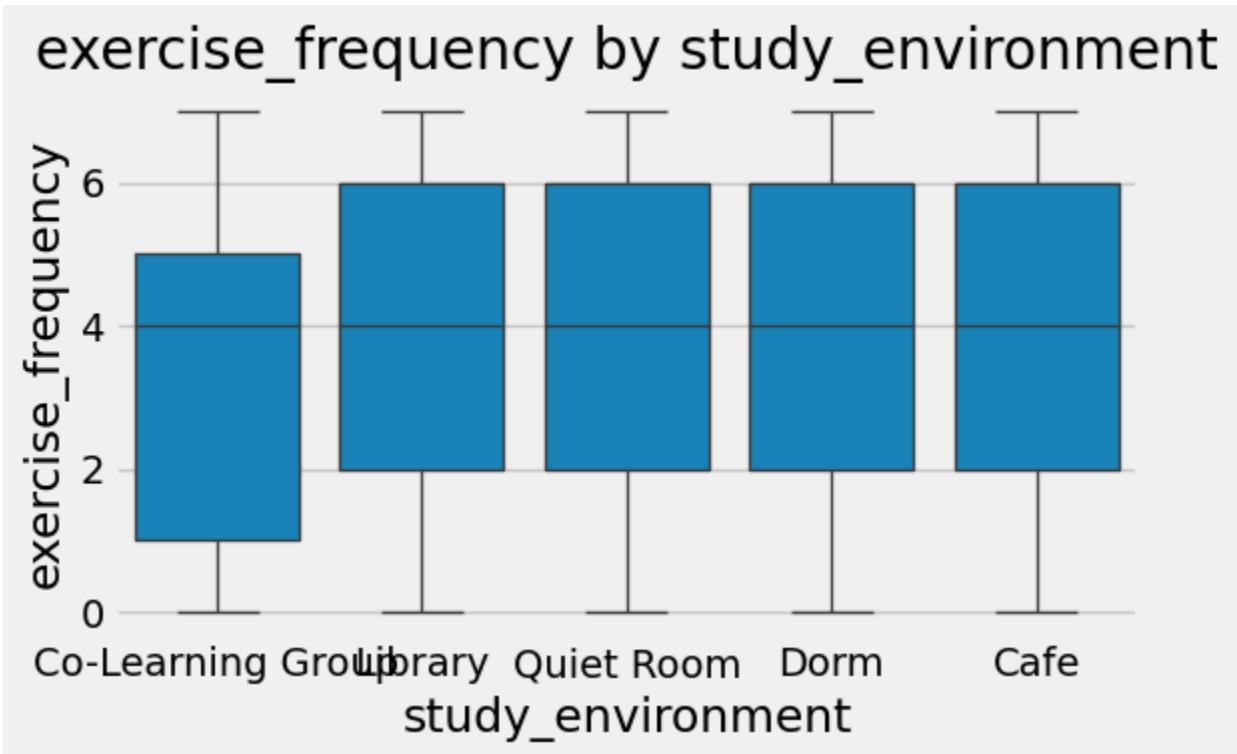
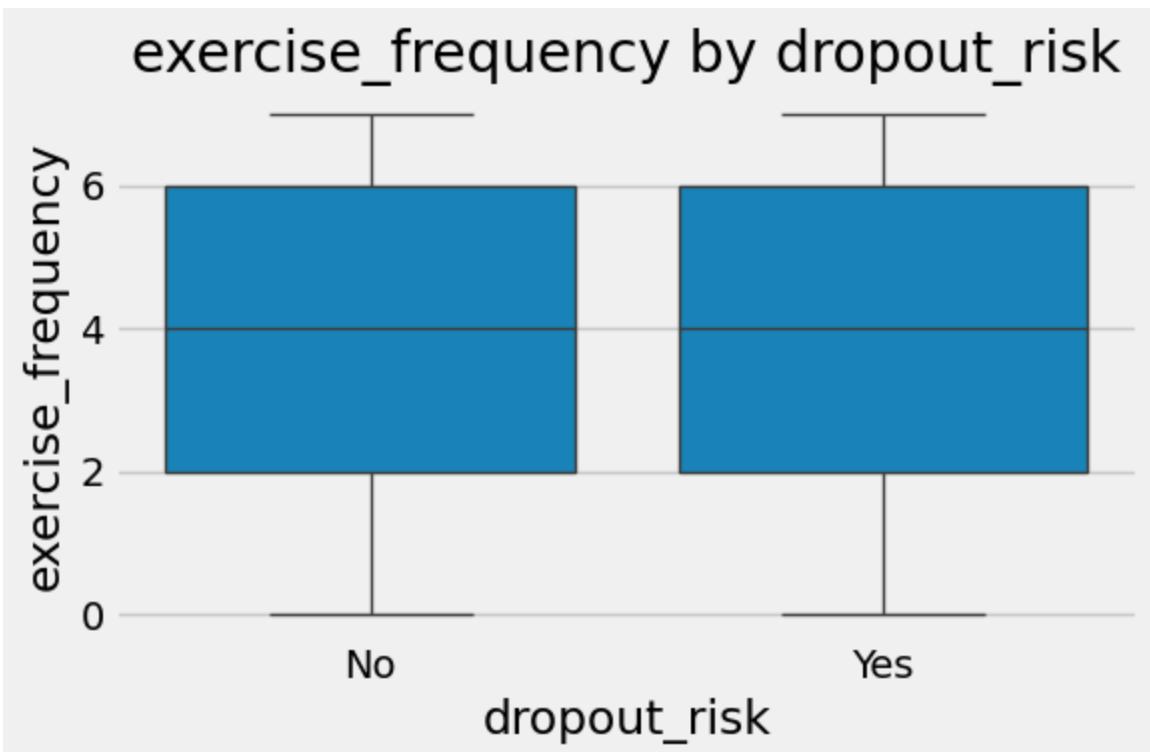


exercise_frequency by internet_quality

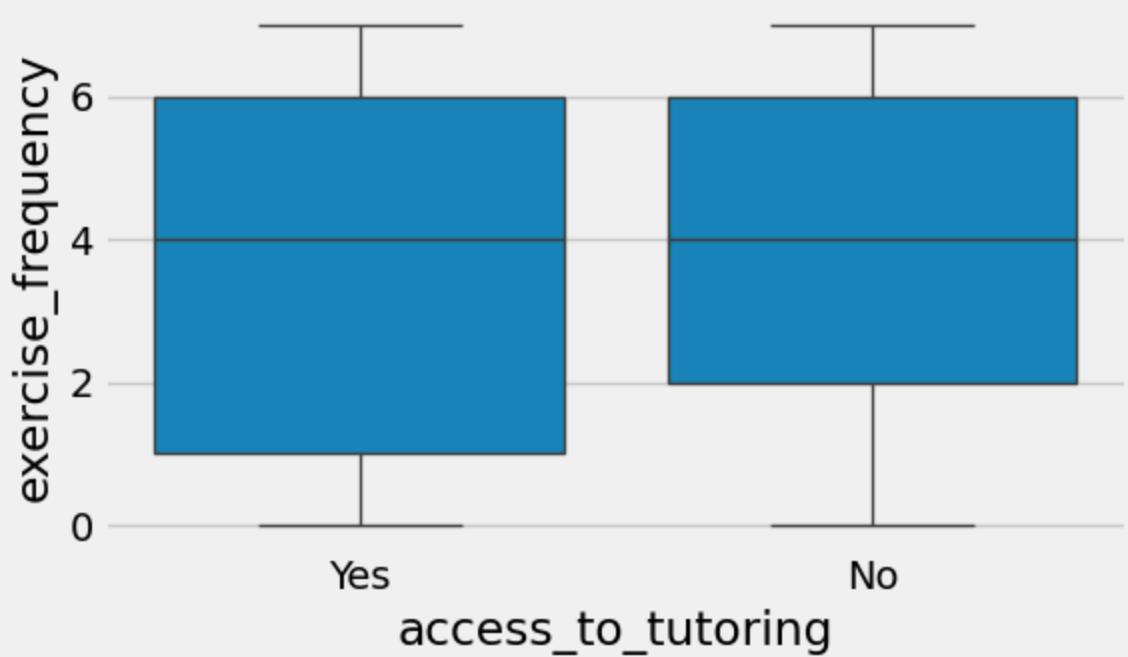


exercise_frequency by extracurricular_participation

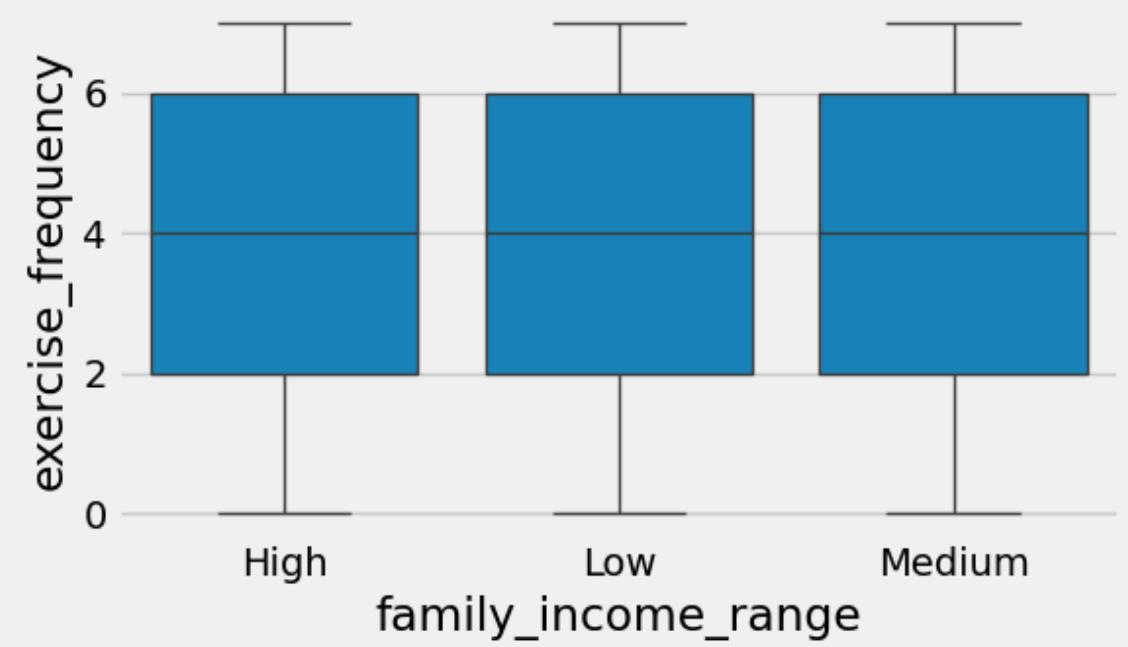


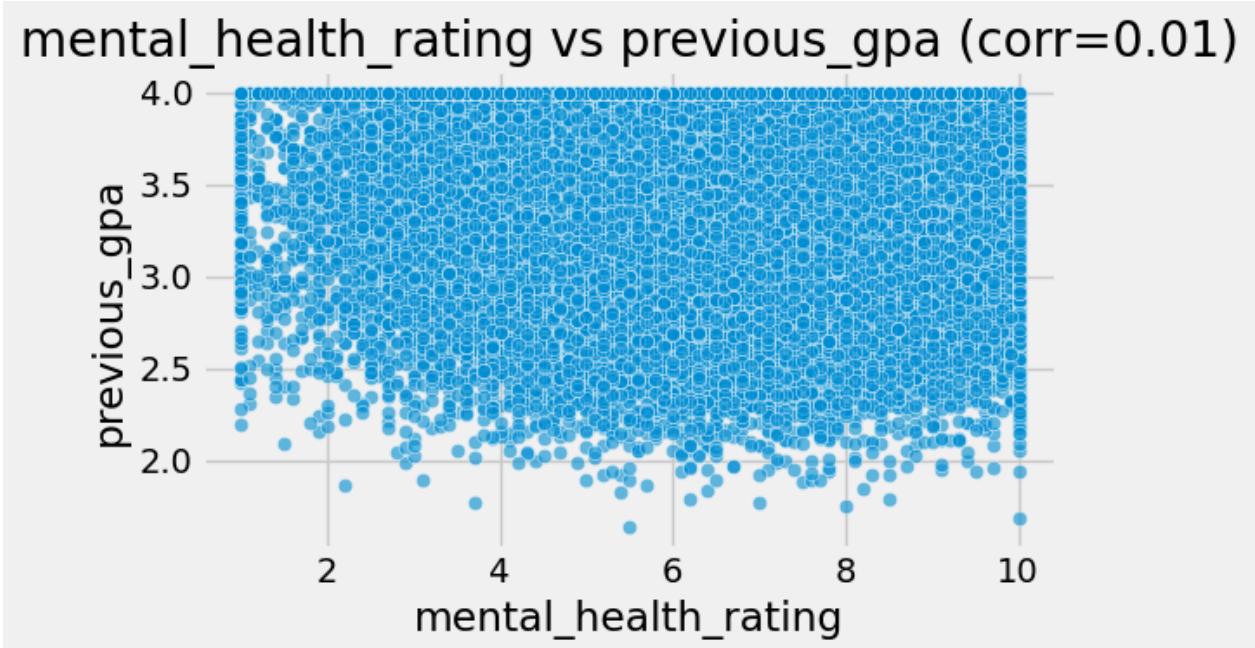
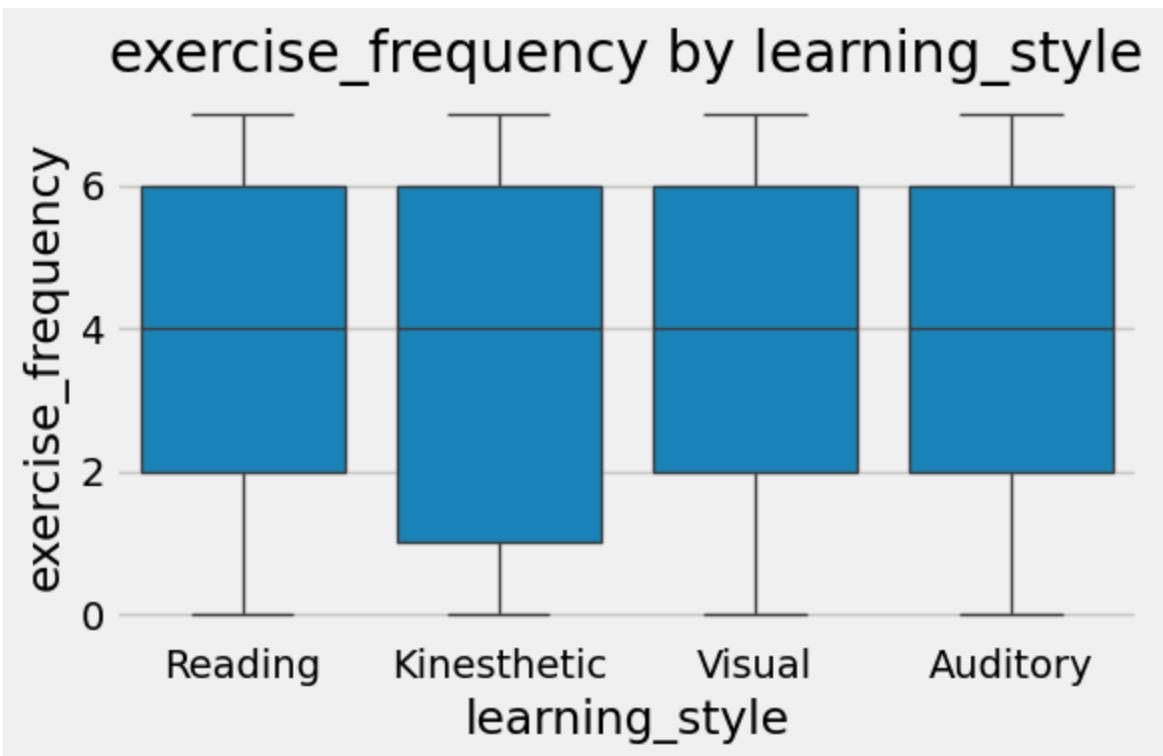


exercise_frequency by access_to_tutoring

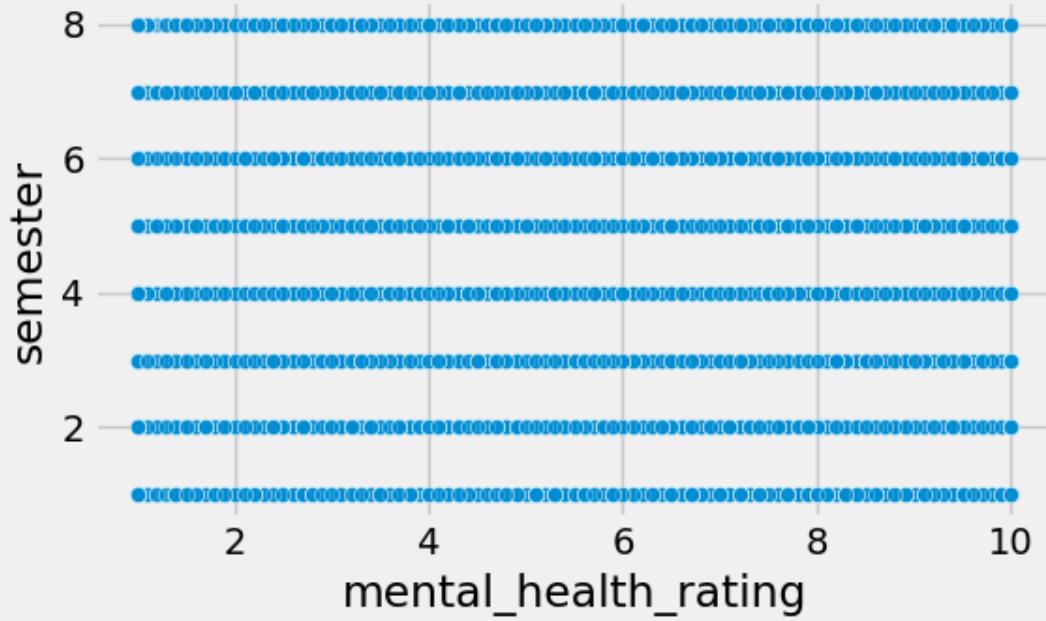


exercise_frequency by family_income_range

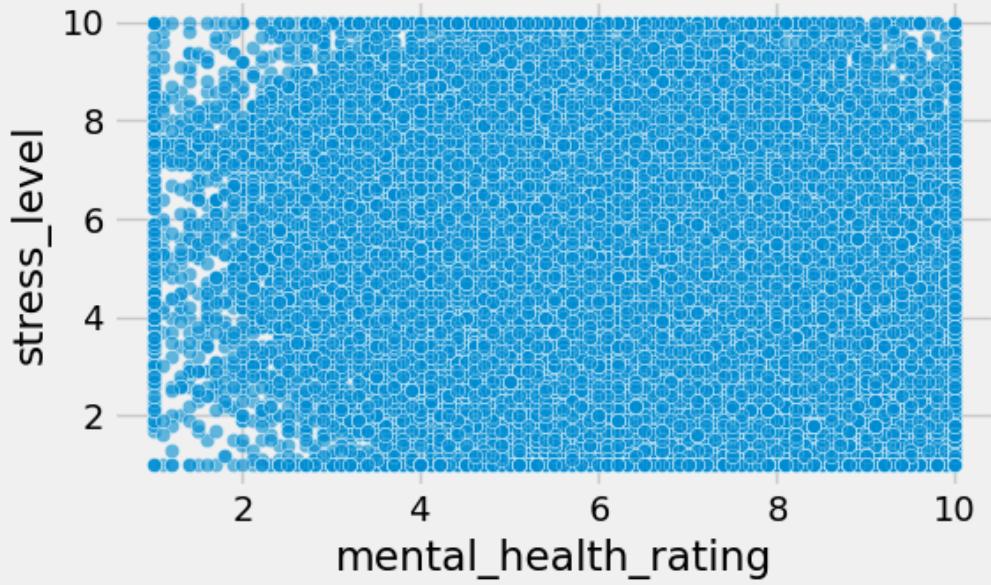




mental_health_rating vs semester (corr=0.00)



mental_health_rating vs stress_level (corr=-0.12)



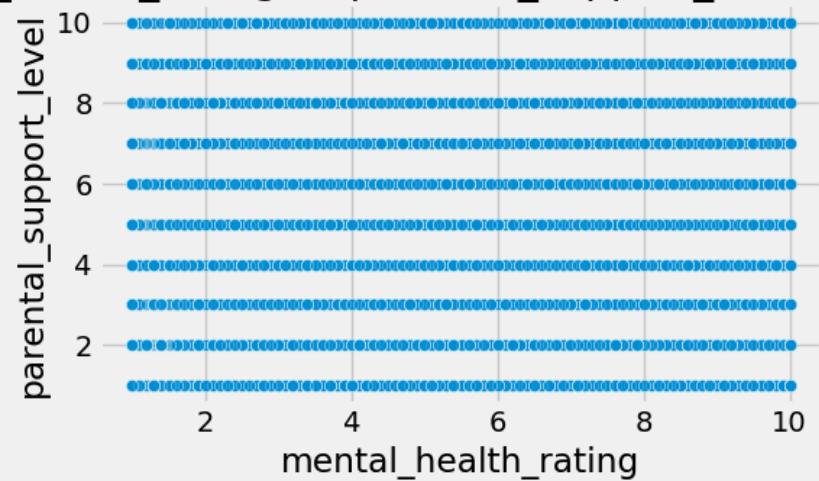
mental_health_rating vs social_activity (corr=0.00)



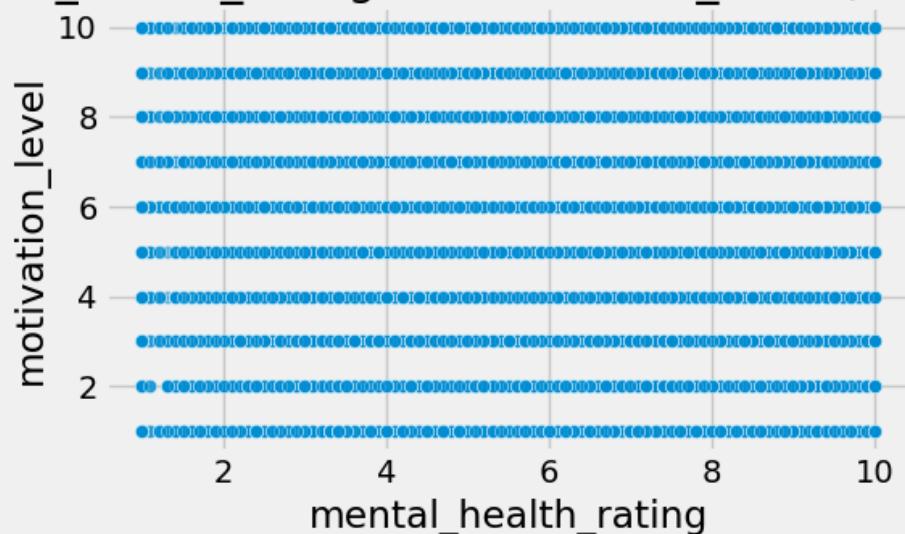
mental_health_rating vs screen_time (corr=0.00)



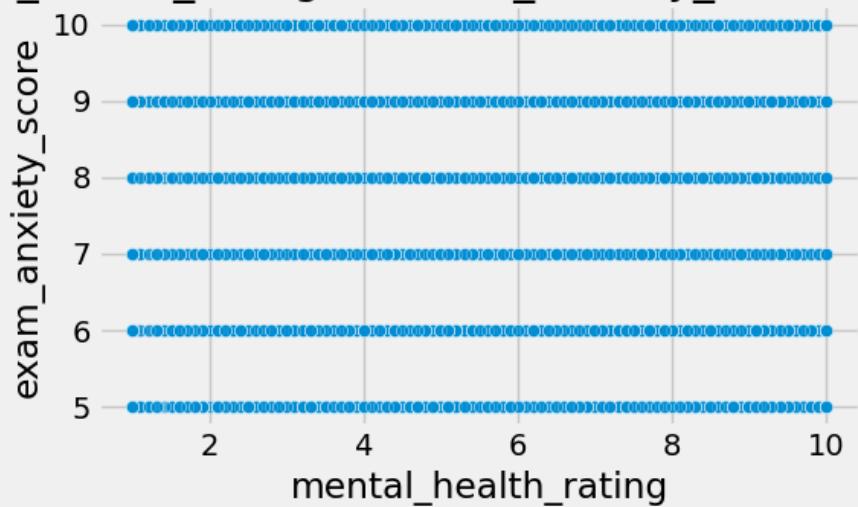
mental_health_rating vs parental_support_level (corr=-0.00)



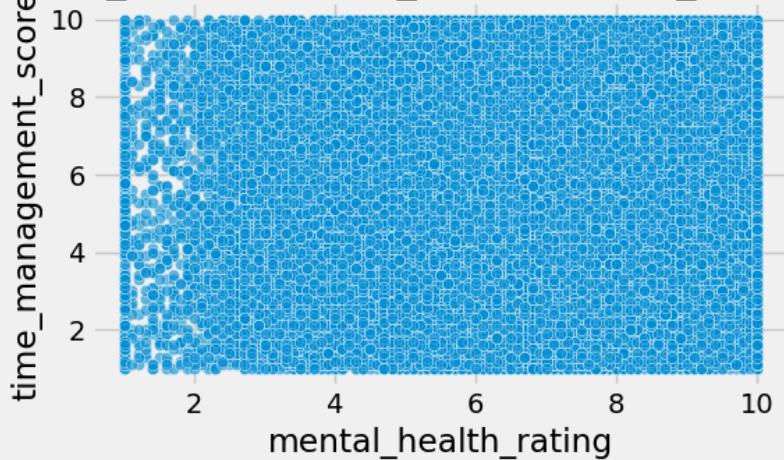
mental_health_rating vs motivation_level (corr=-0.00)



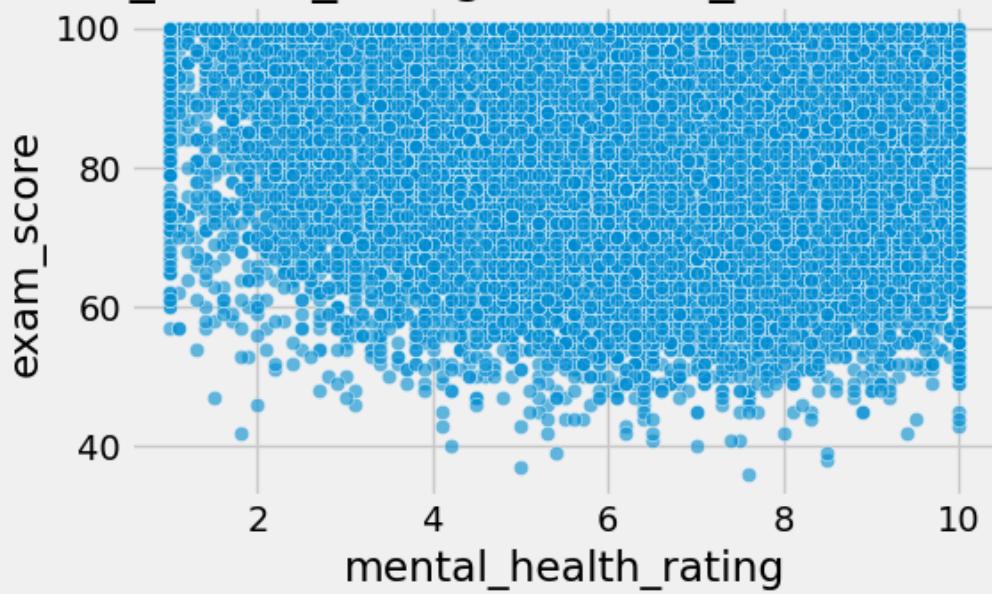
mental_health_rating vs exam_anxiety_score (corr=0.00)

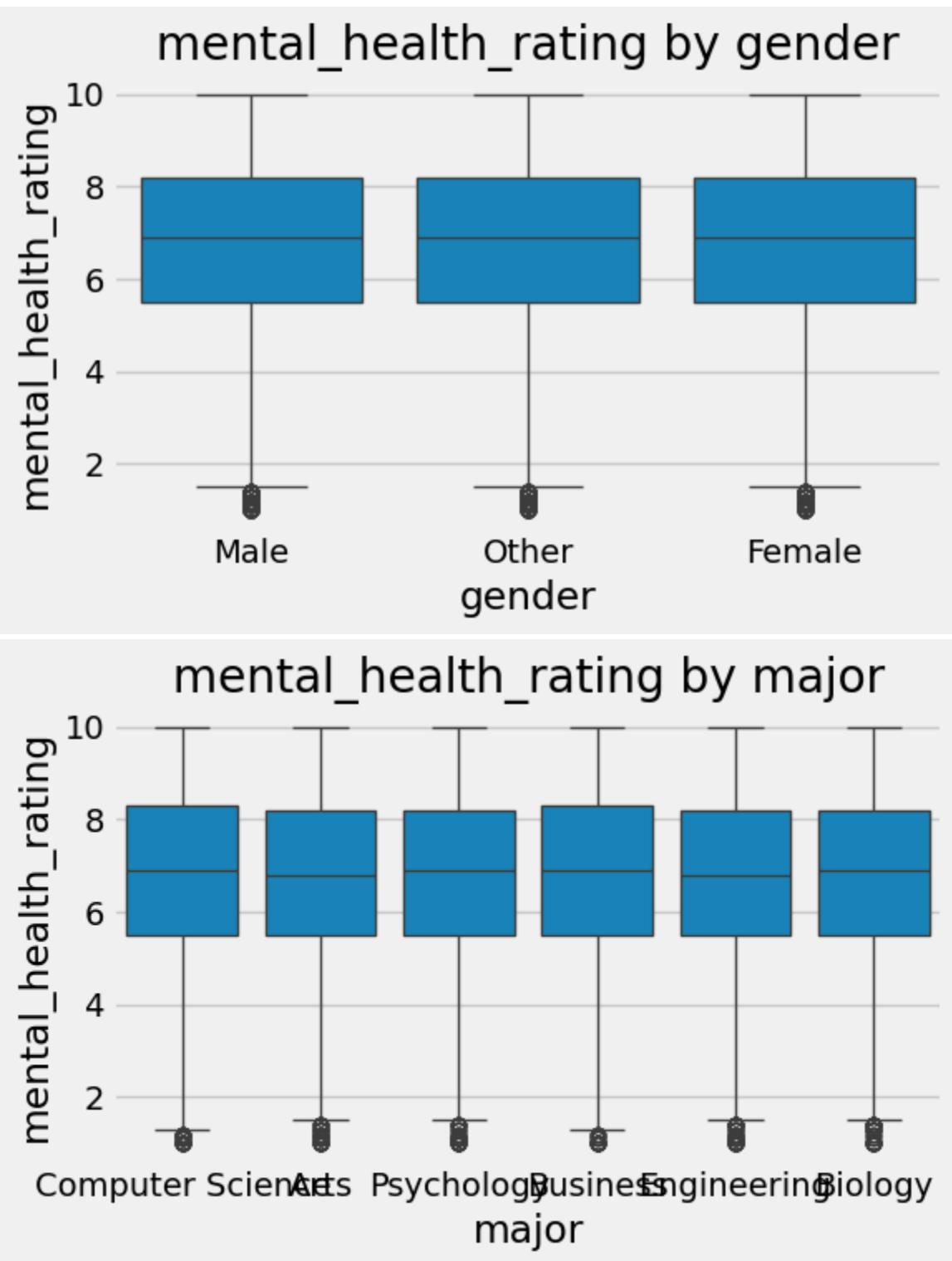


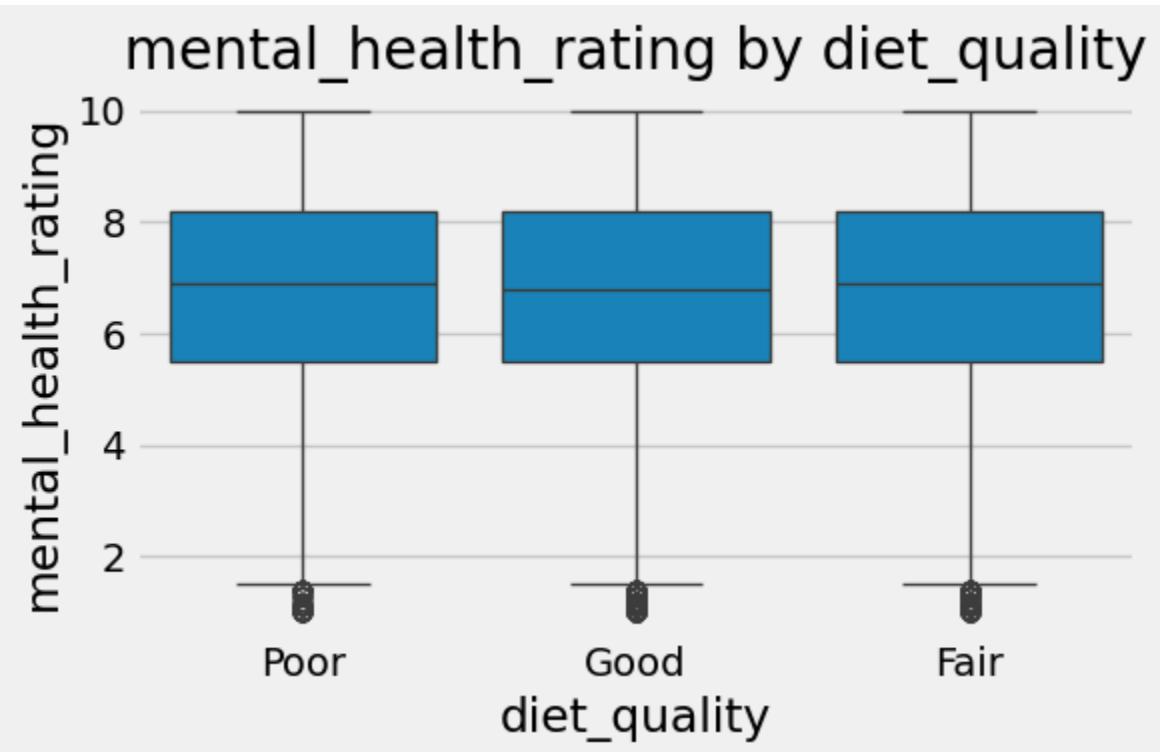
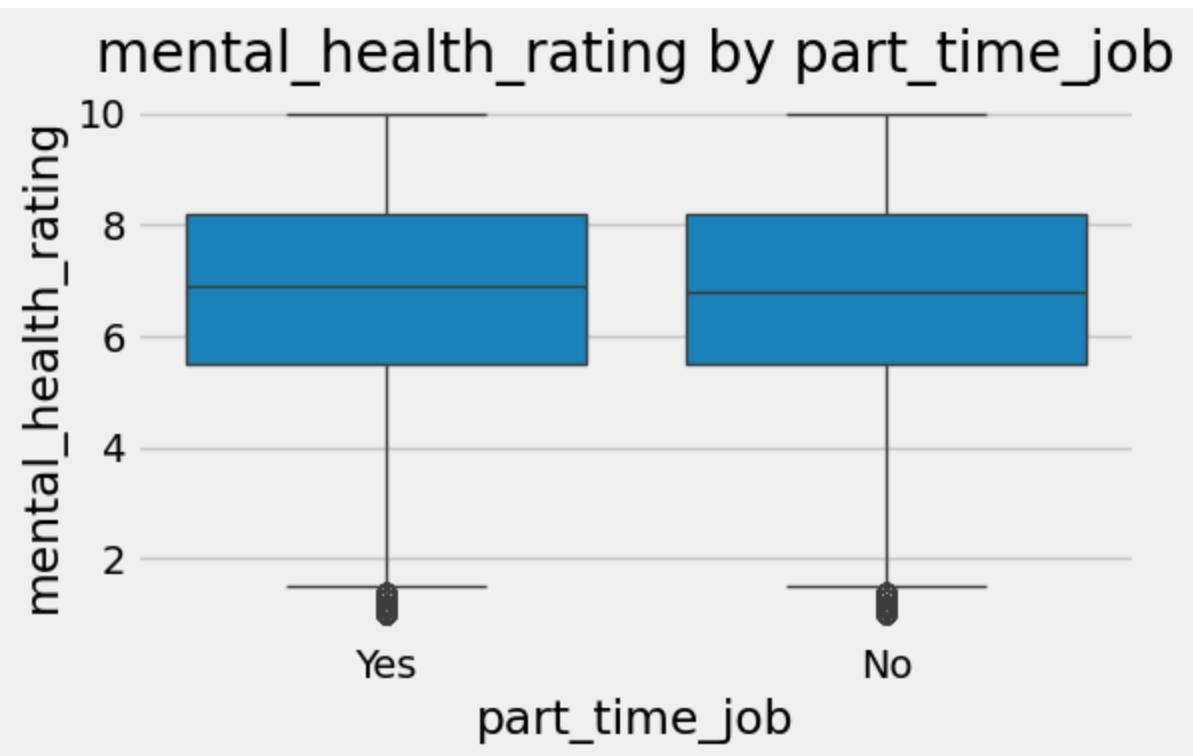
mental_health_rating vs time_management_score (corr=-0.00)

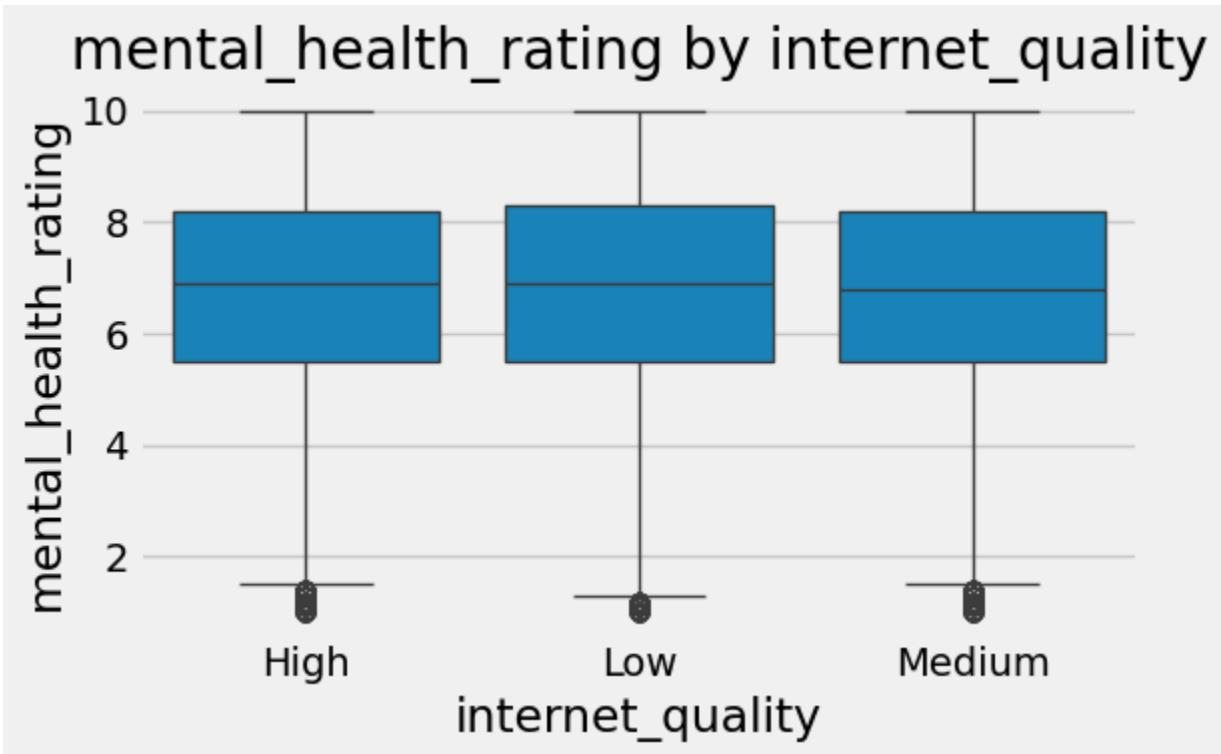
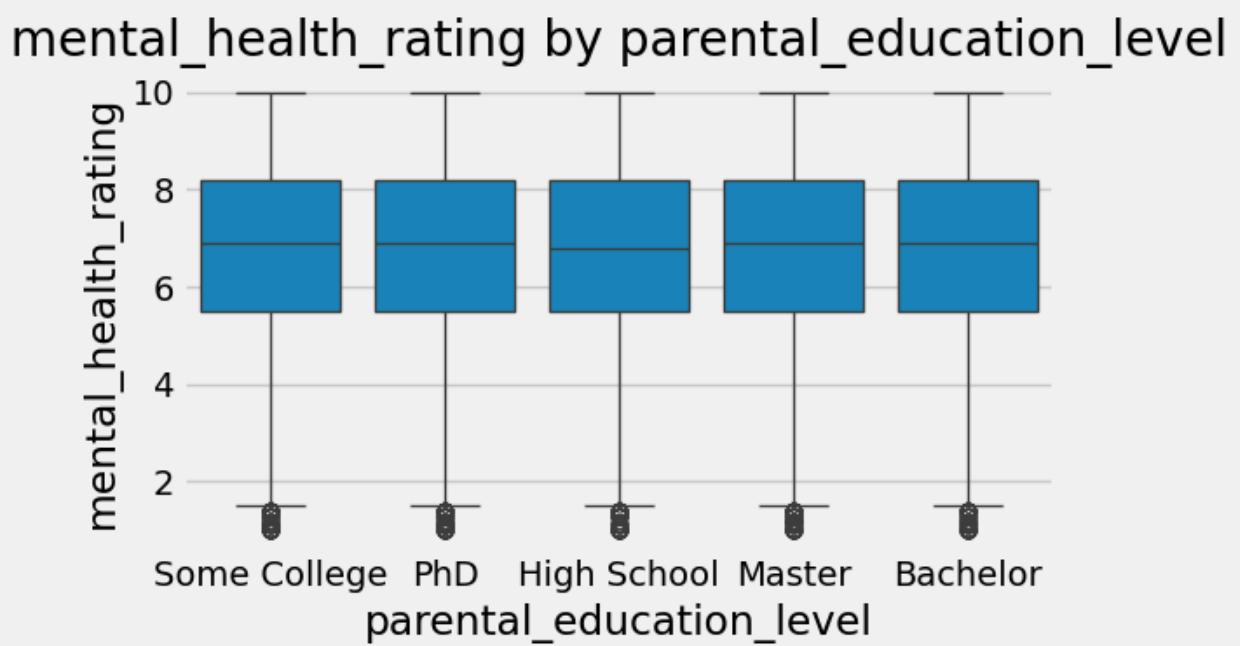


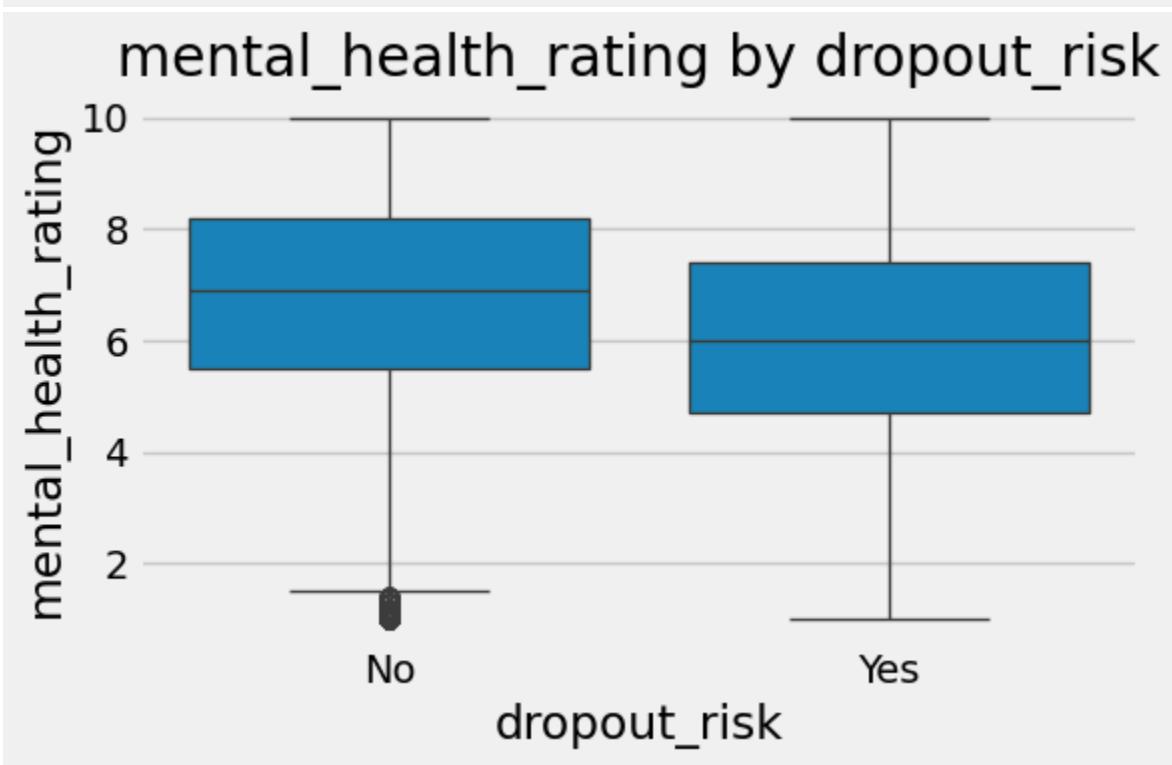
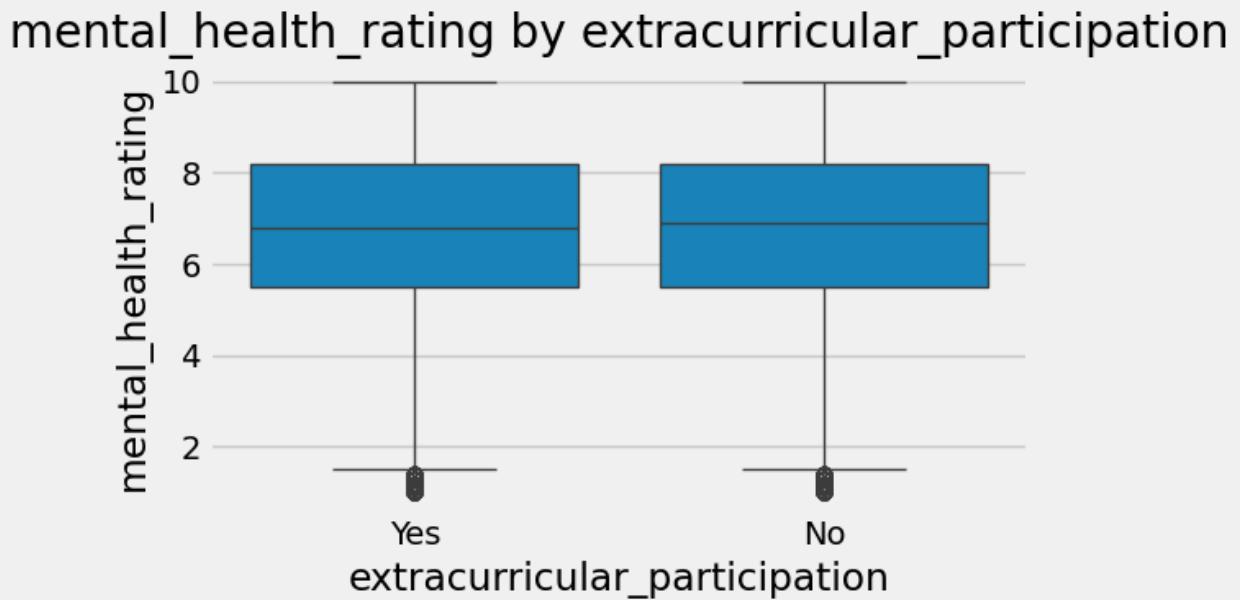
mental_health_rating vs exam_score (corr=0.01)

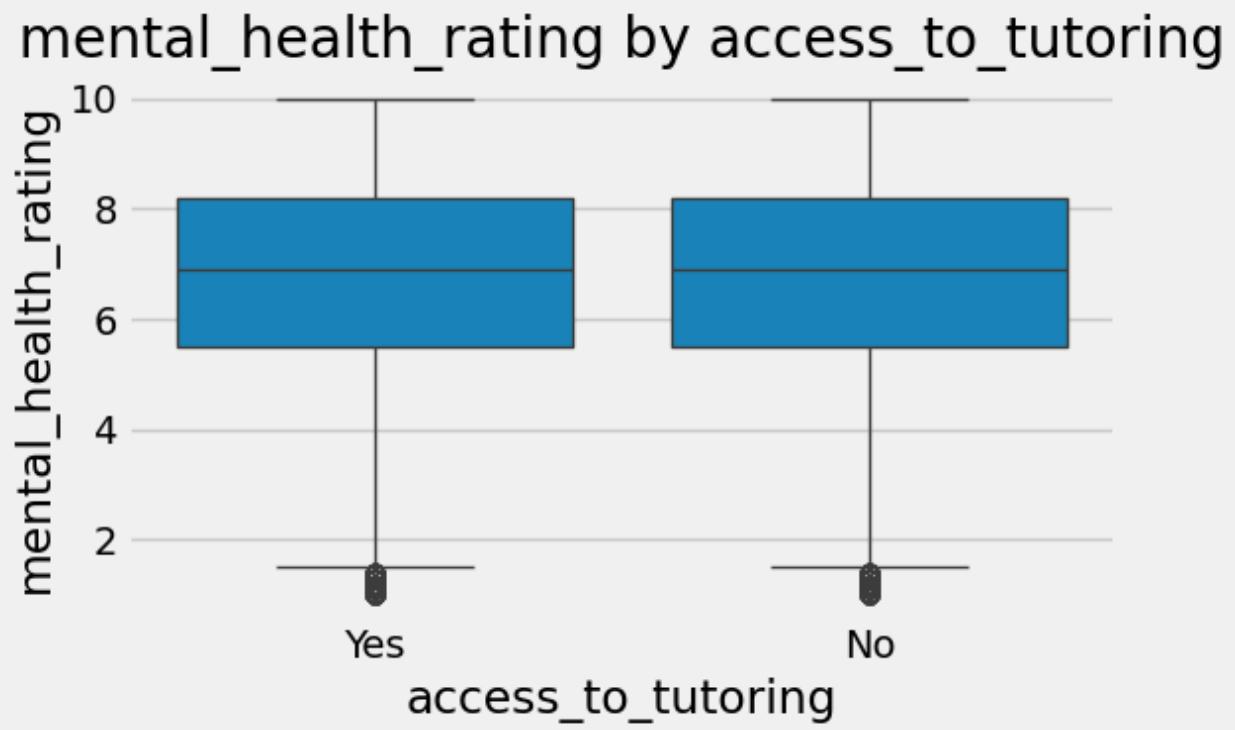
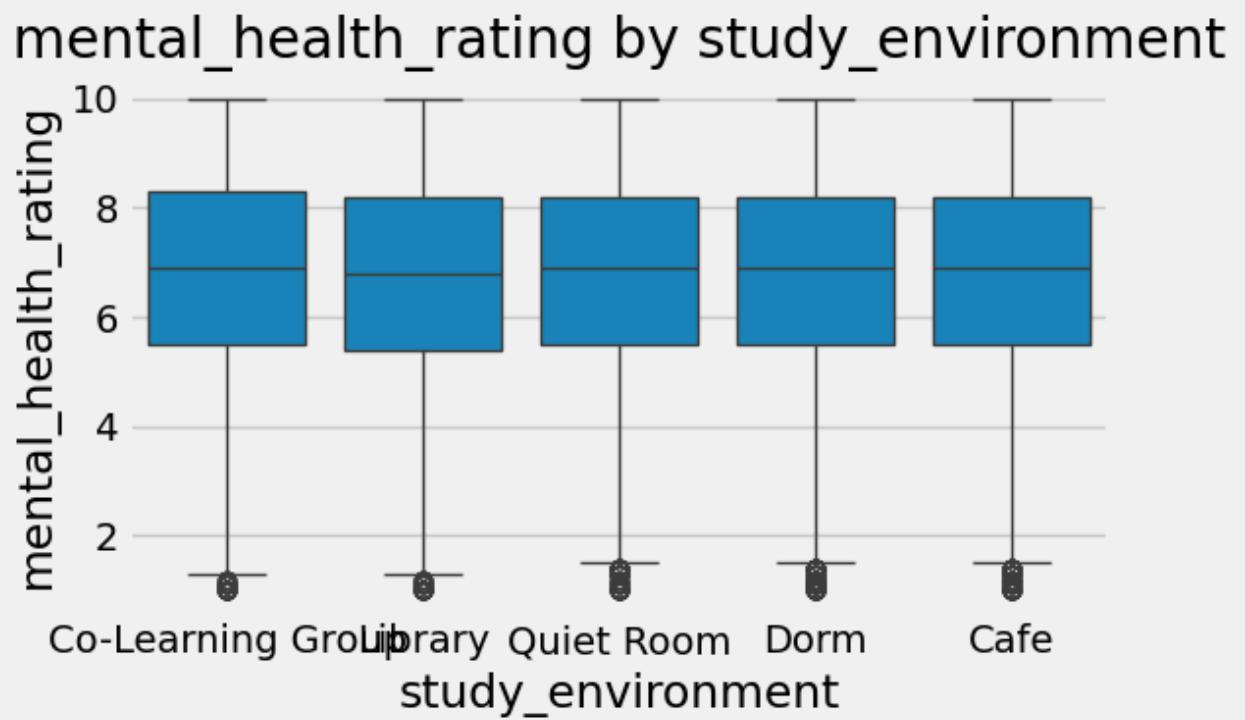


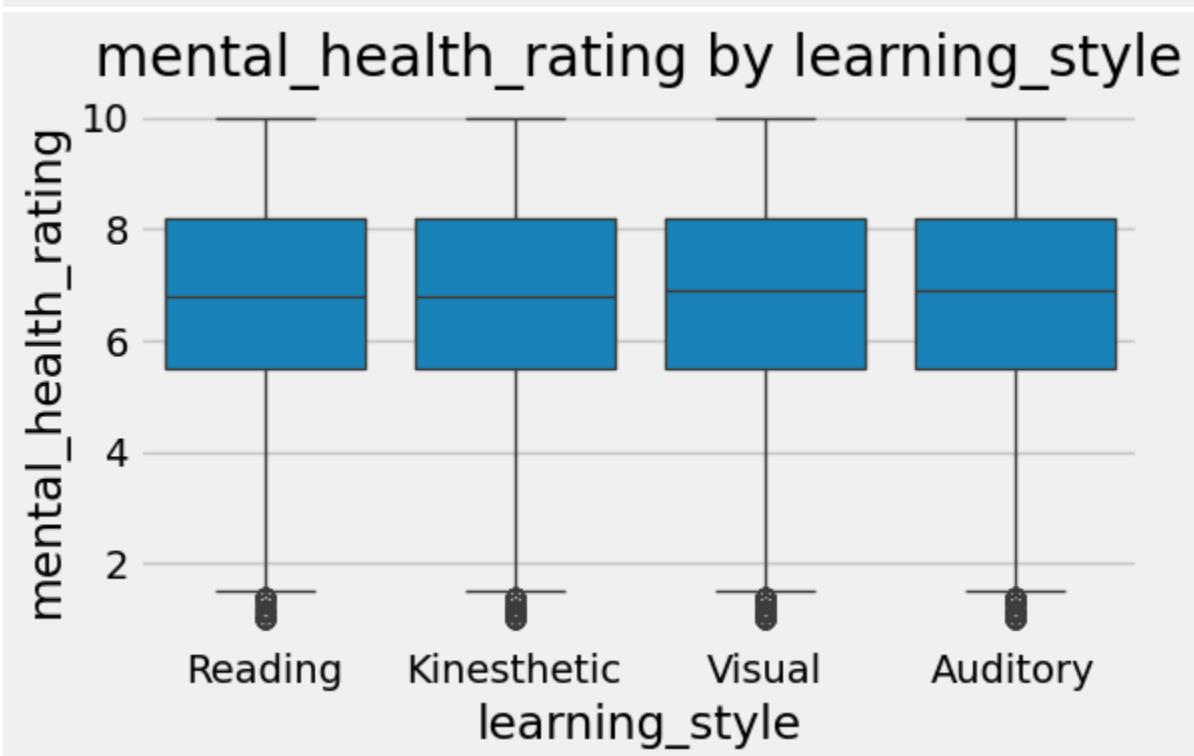
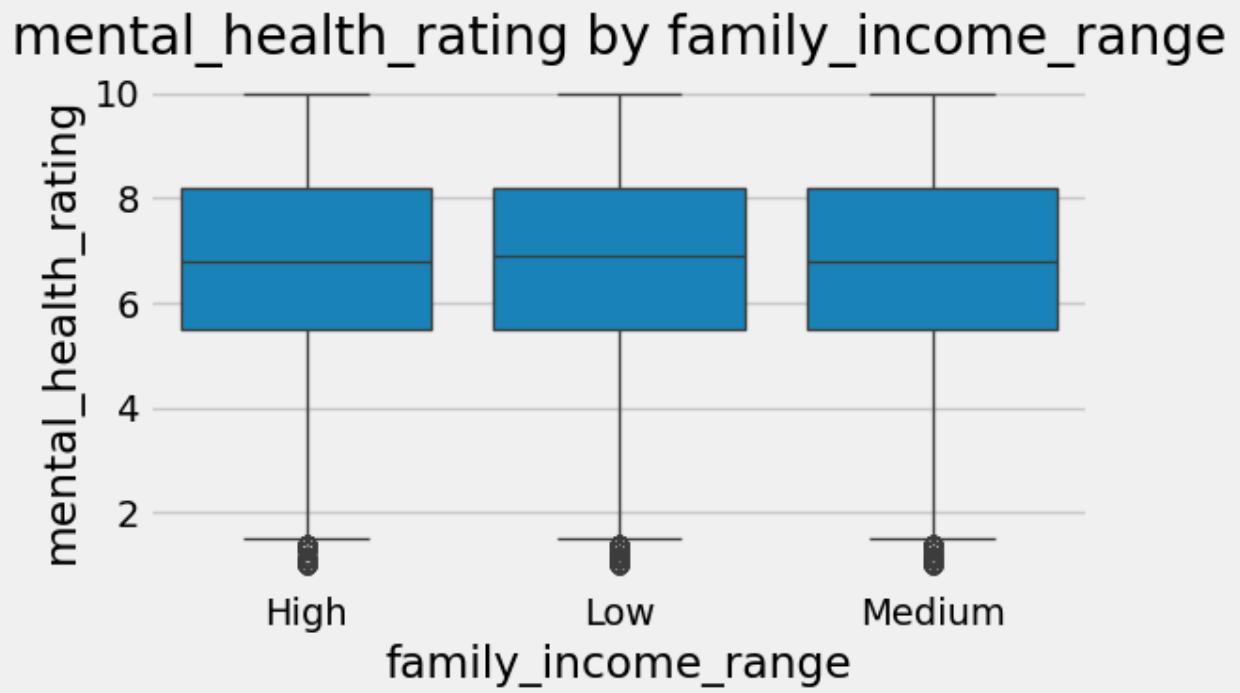


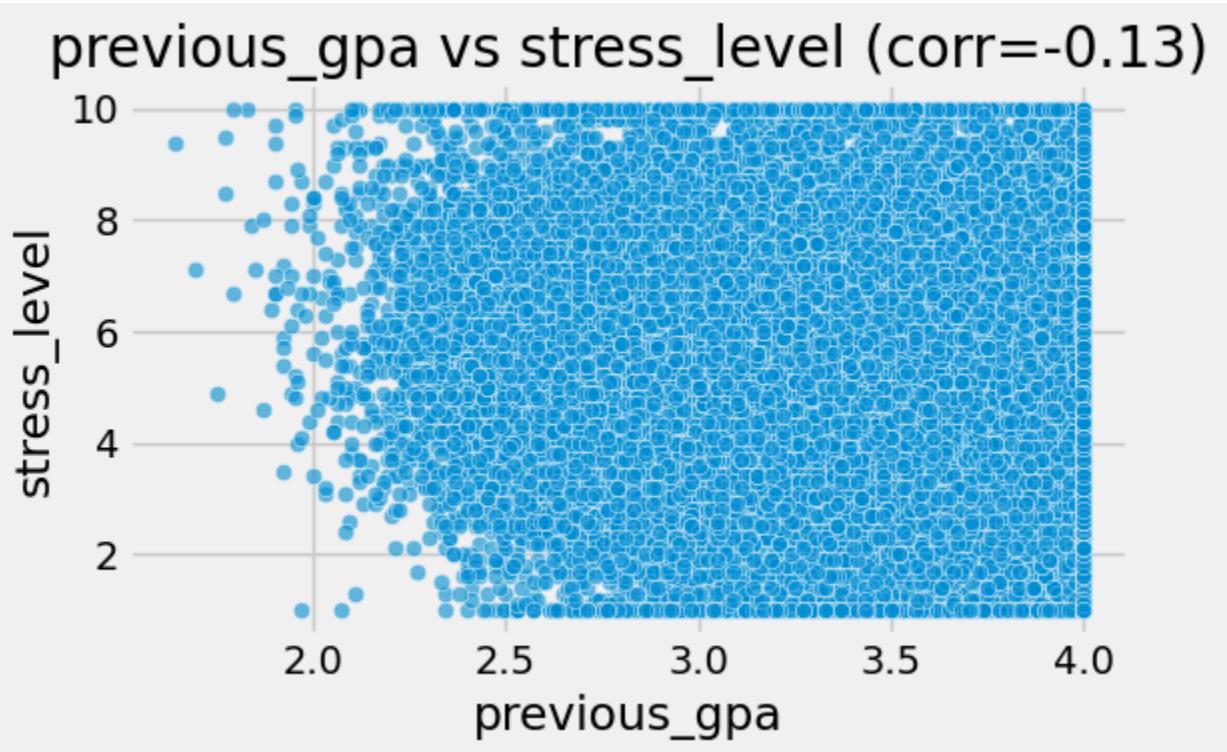
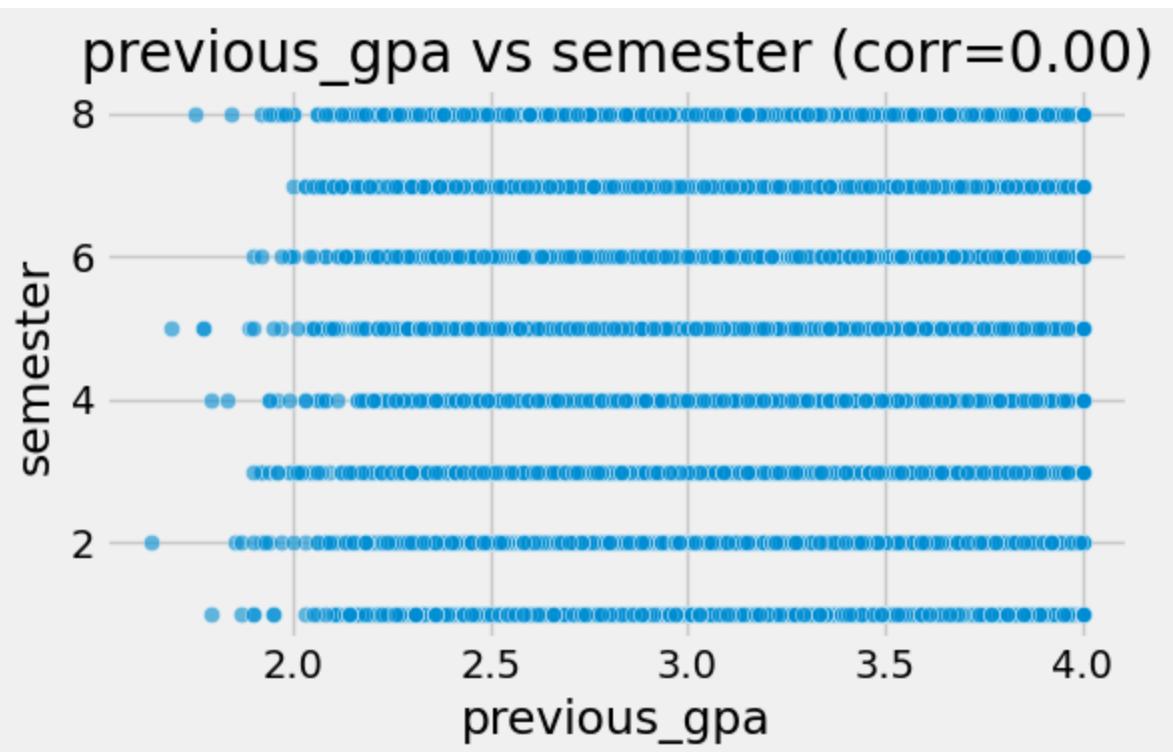




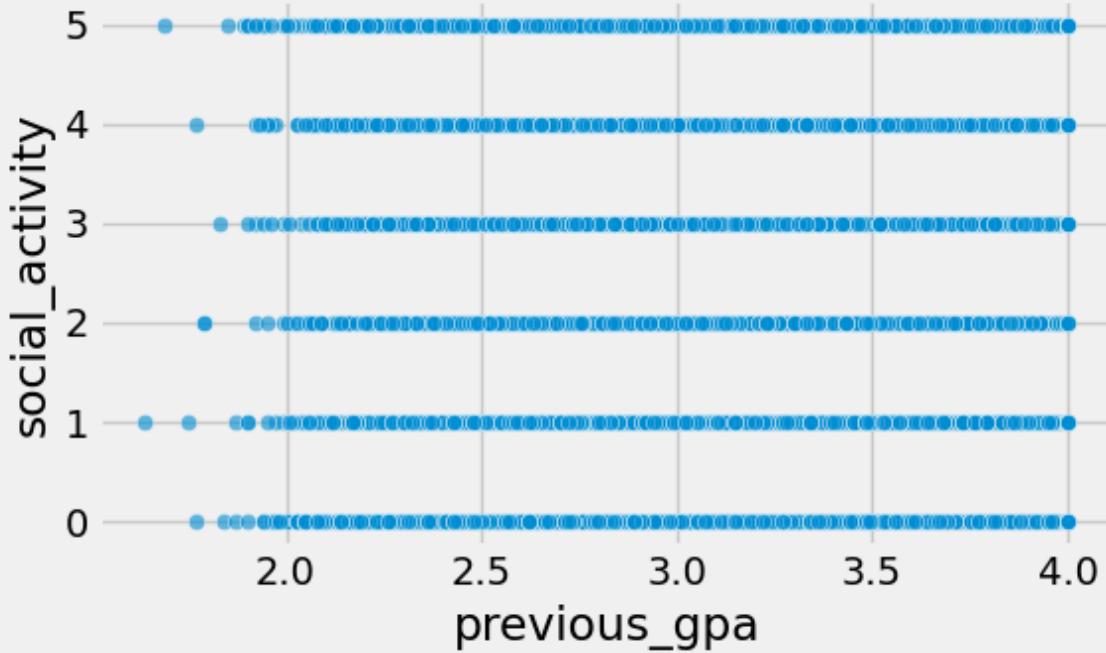




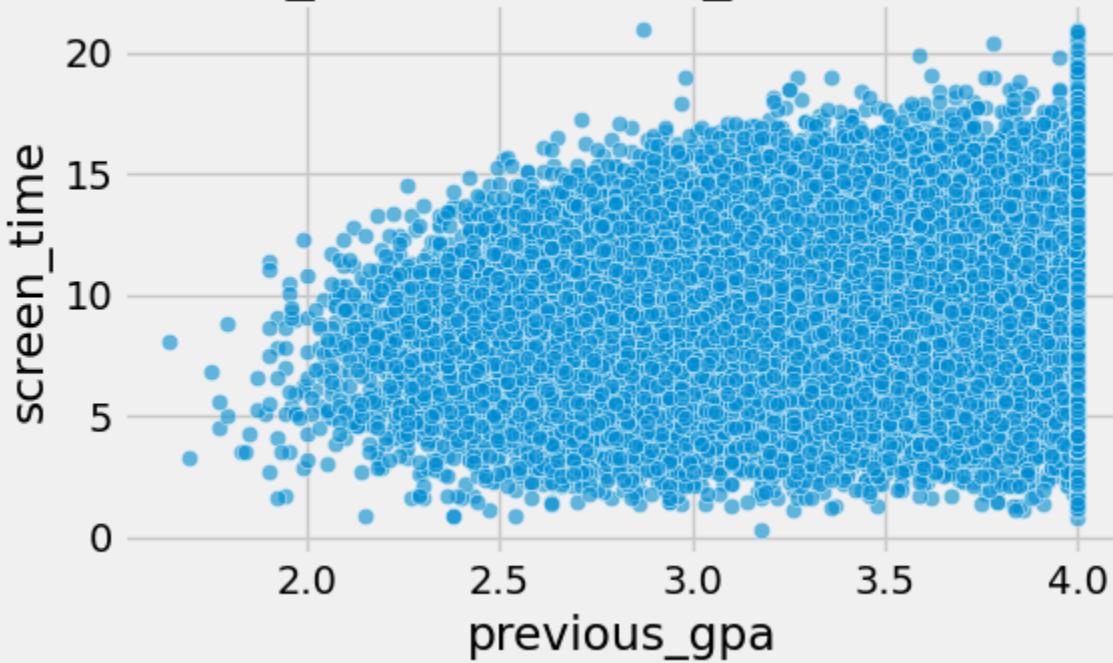




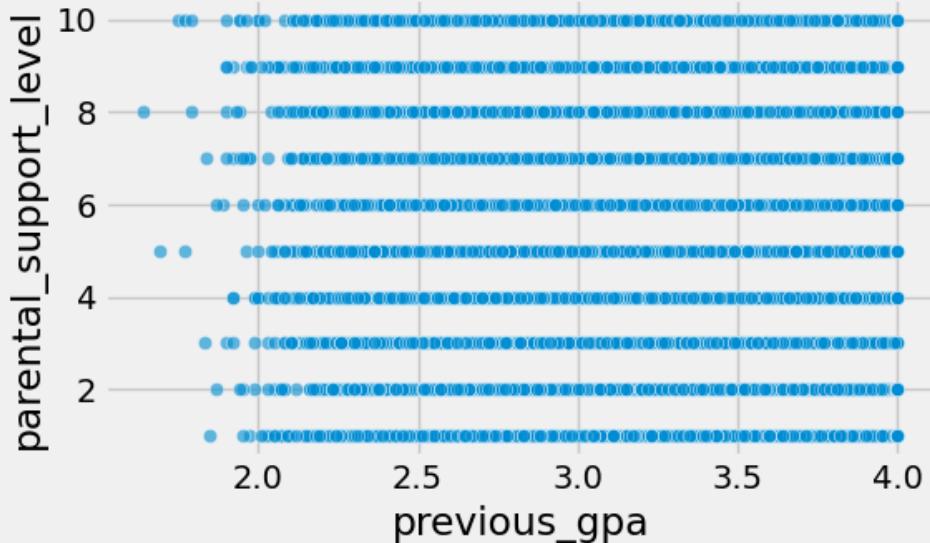
previous_gpa vs social_activity (corr=-0.00)



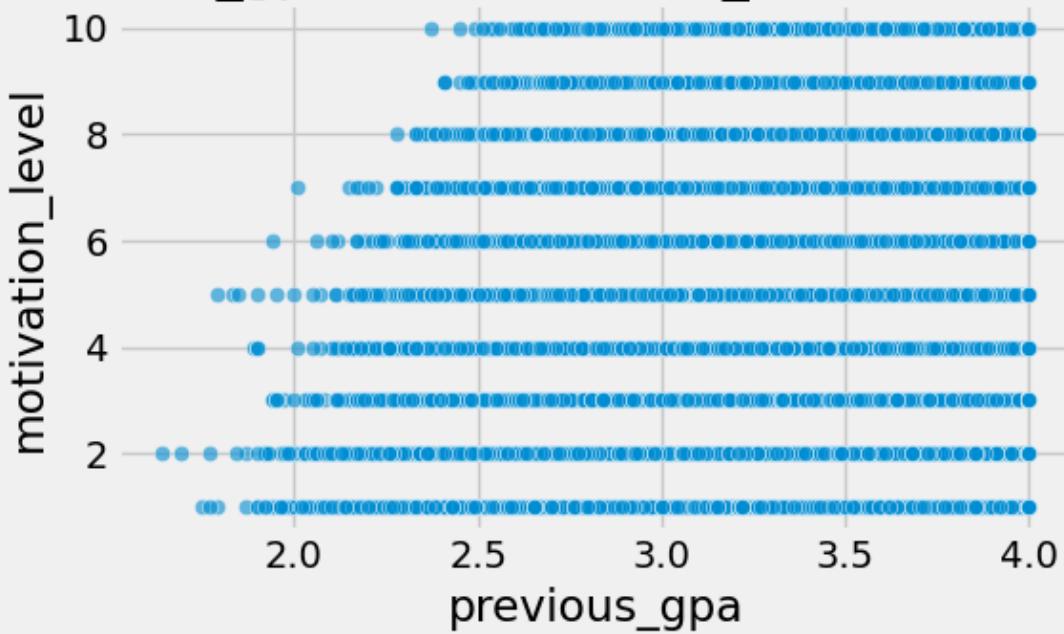
previous_gpa vs screen_time (corr=0.18)



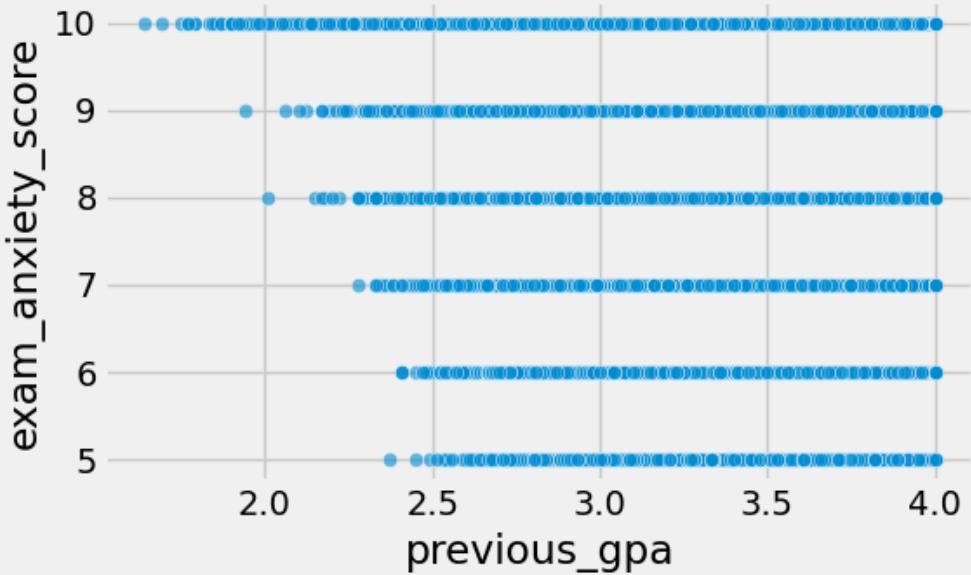
previous_gpa vs parental_support_level (corr=-0.01)



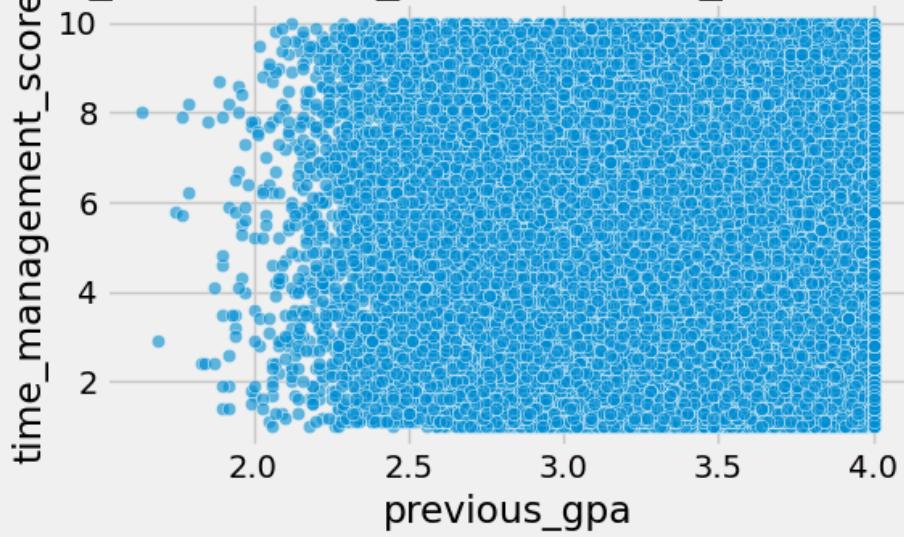
previous_gpa vs motivation_level (corr=0.27)



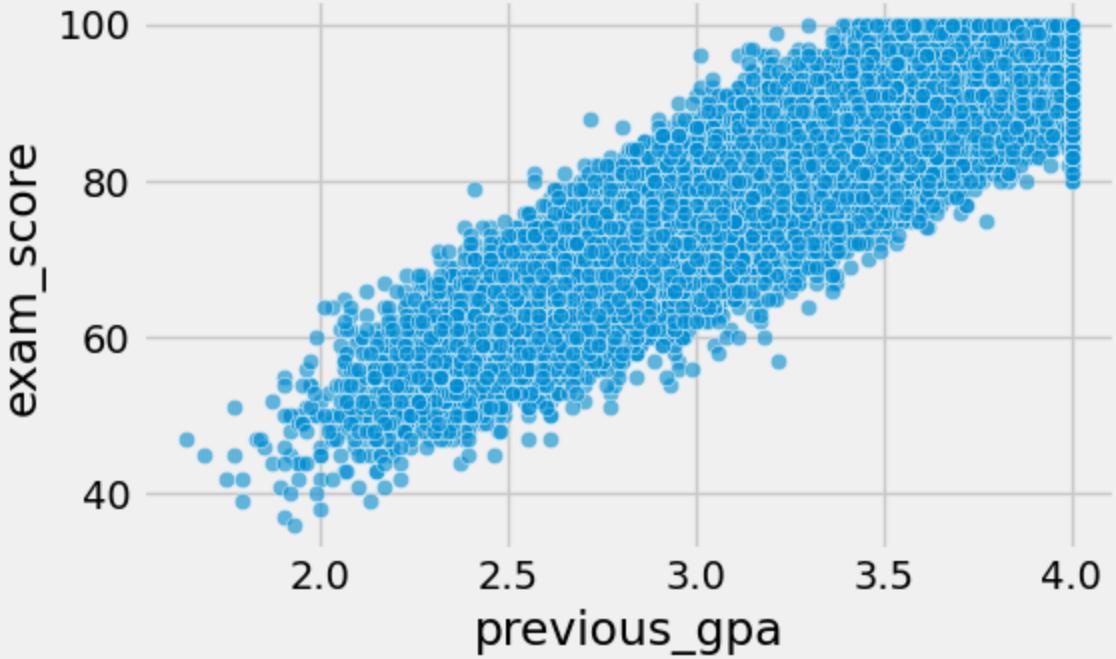
previous_gpa vs exam_anxiety_score (corr=-0.26)



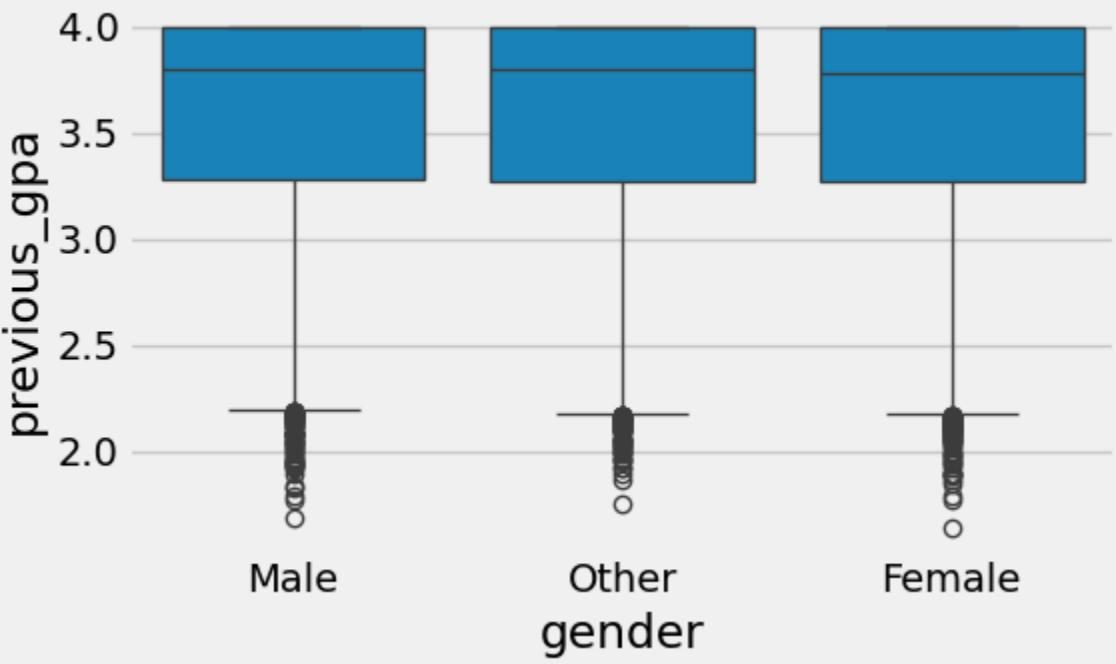
previous_gpa vs time_management_score (corr=0.01)

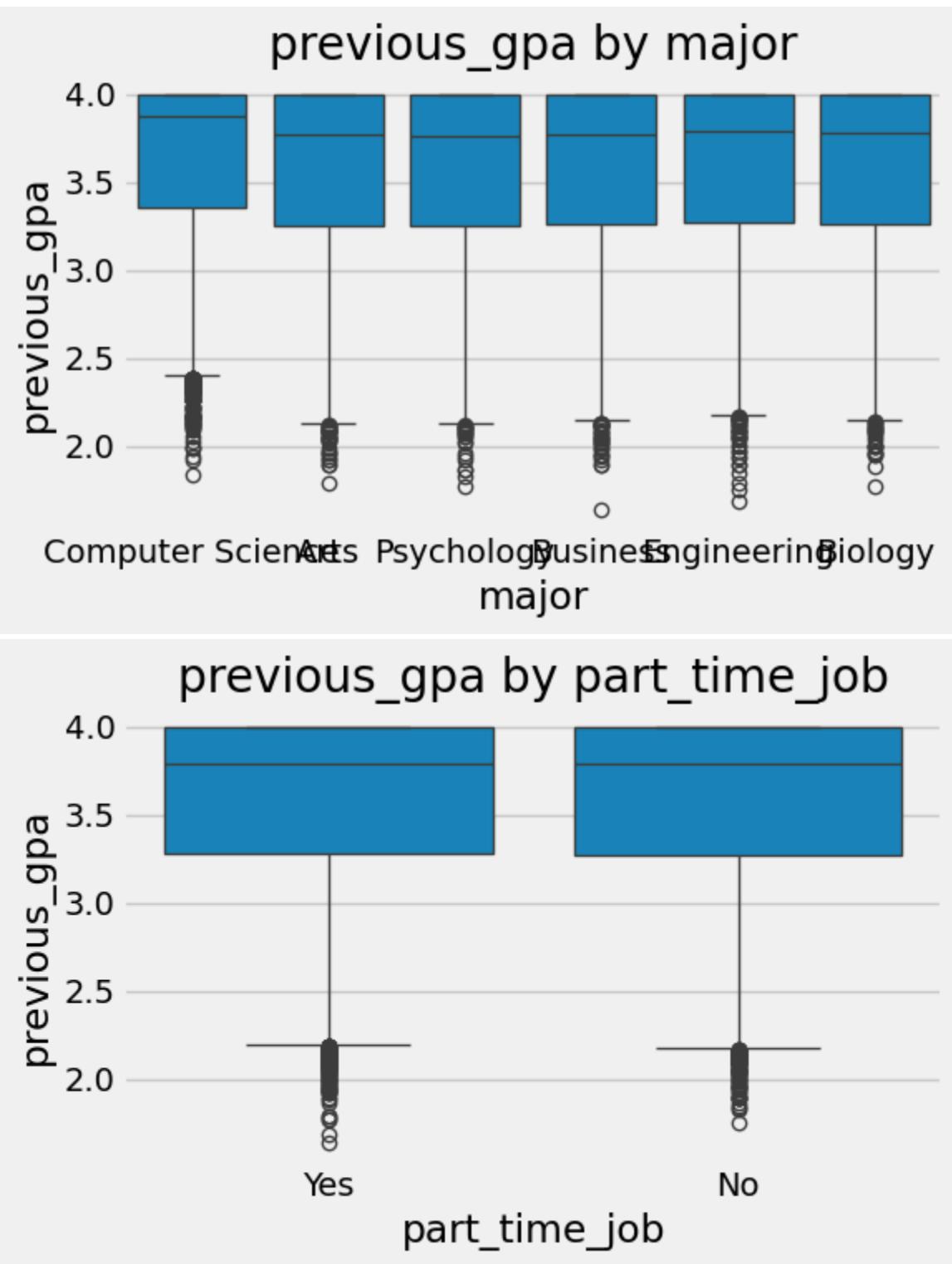


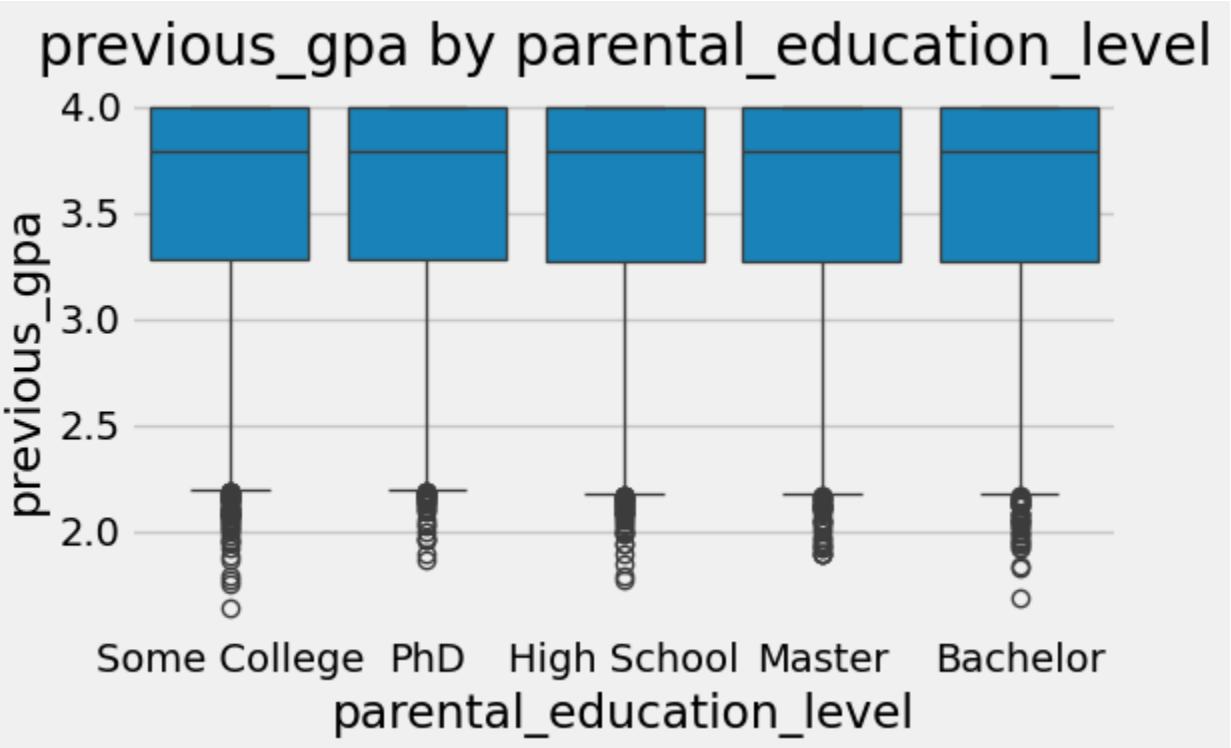
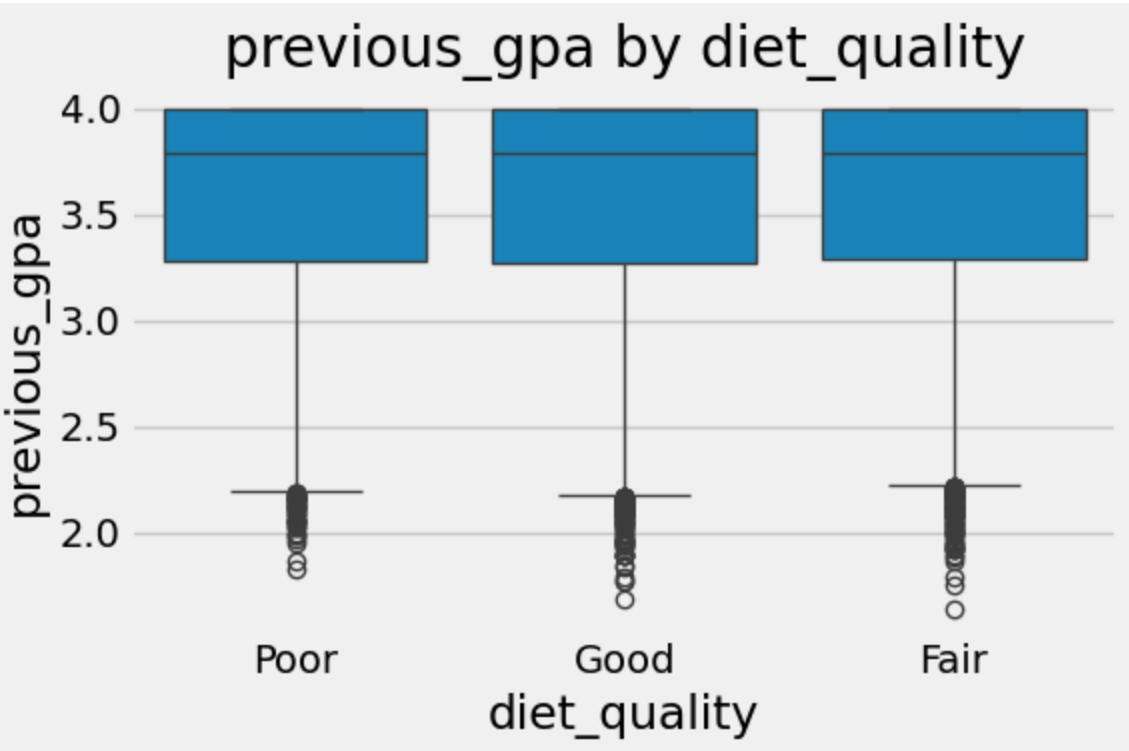
previous_gpa vs exam_score (corr=0.93)

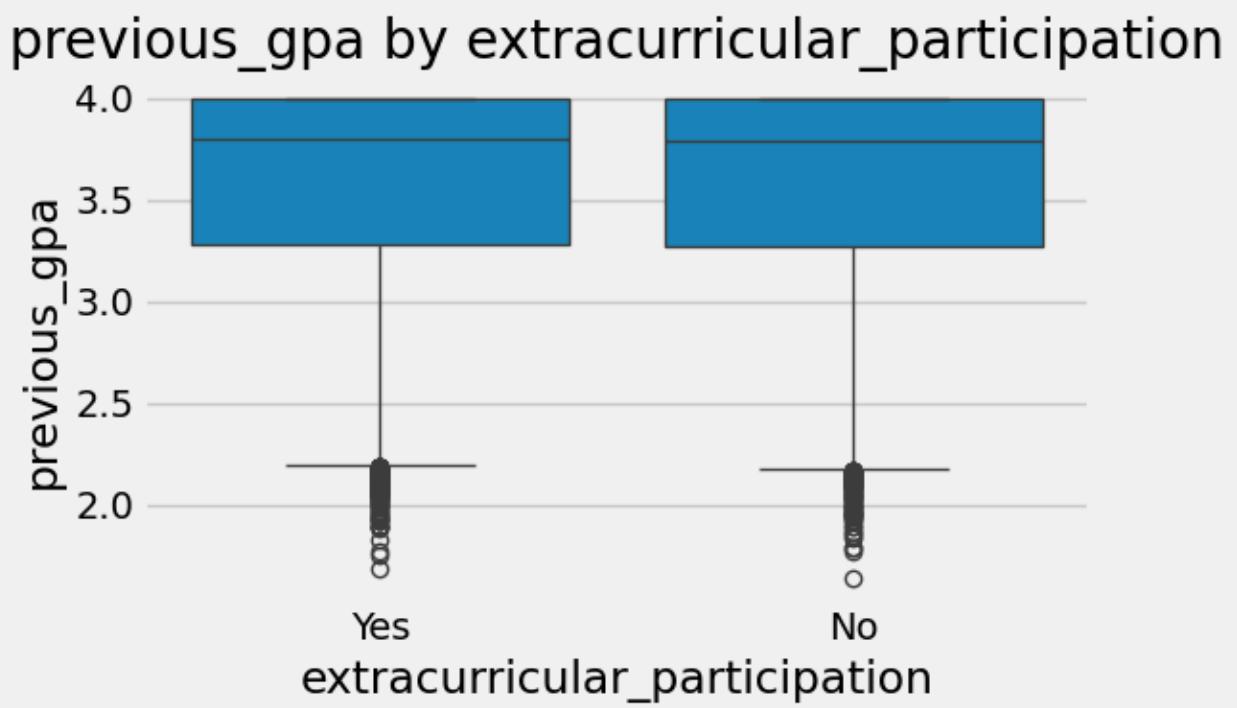
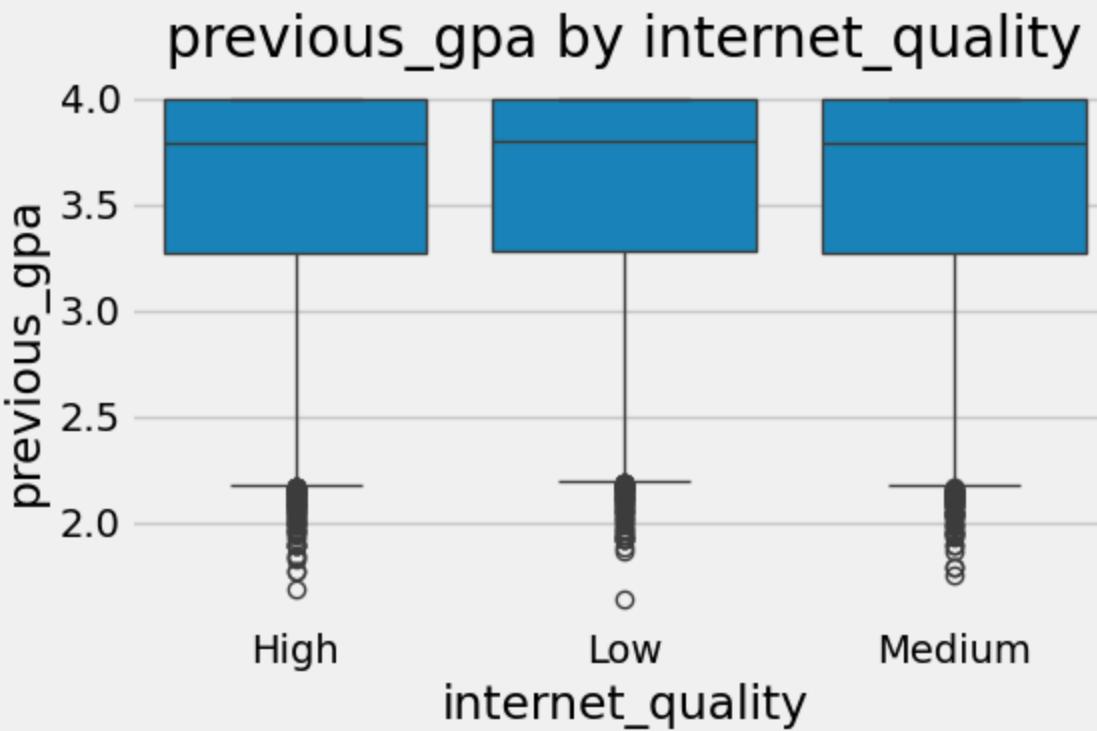


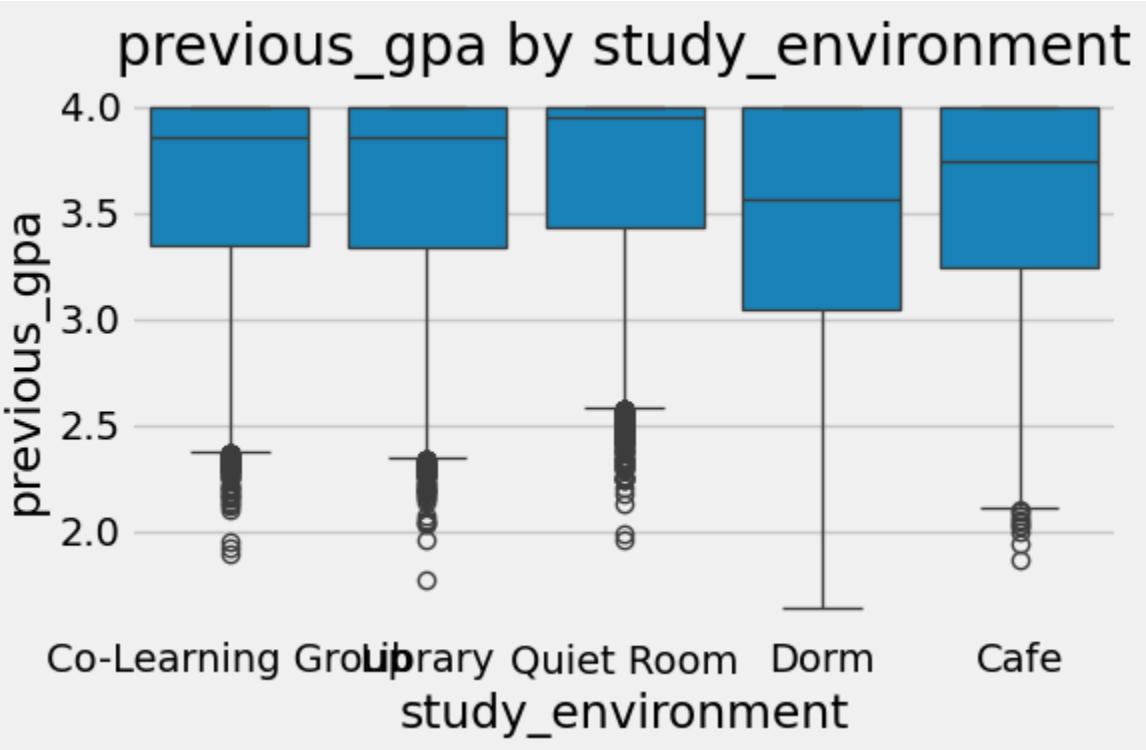
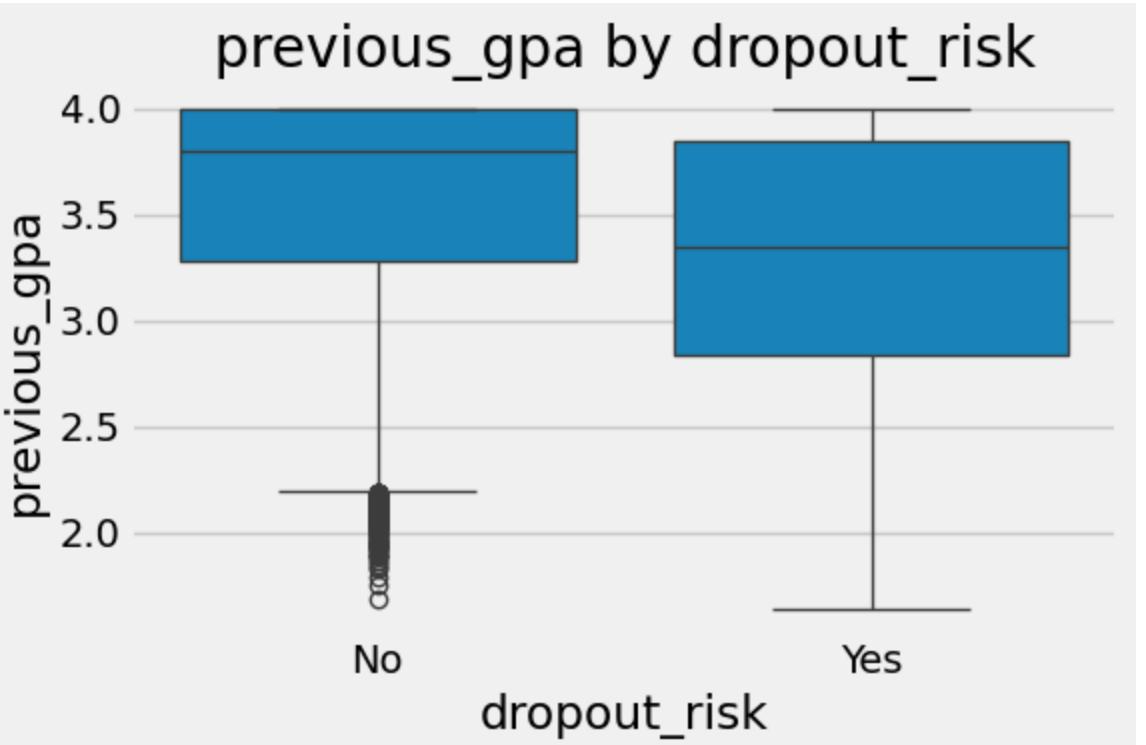
previous_gpa by gender

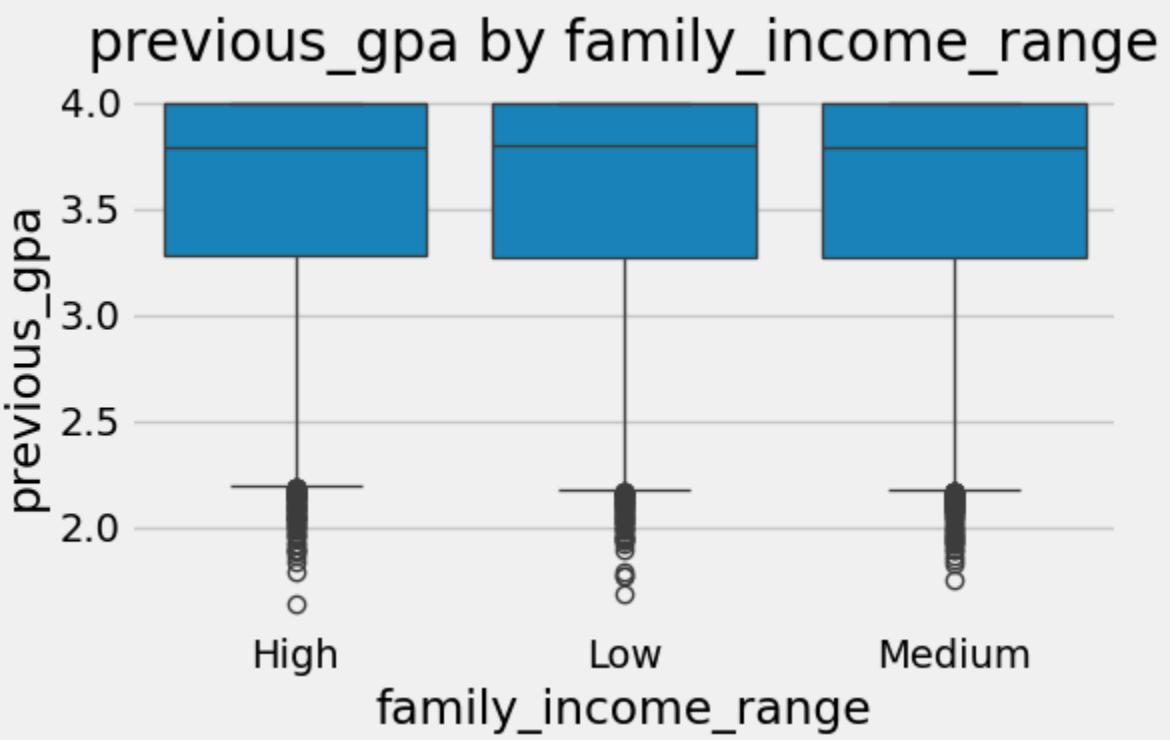
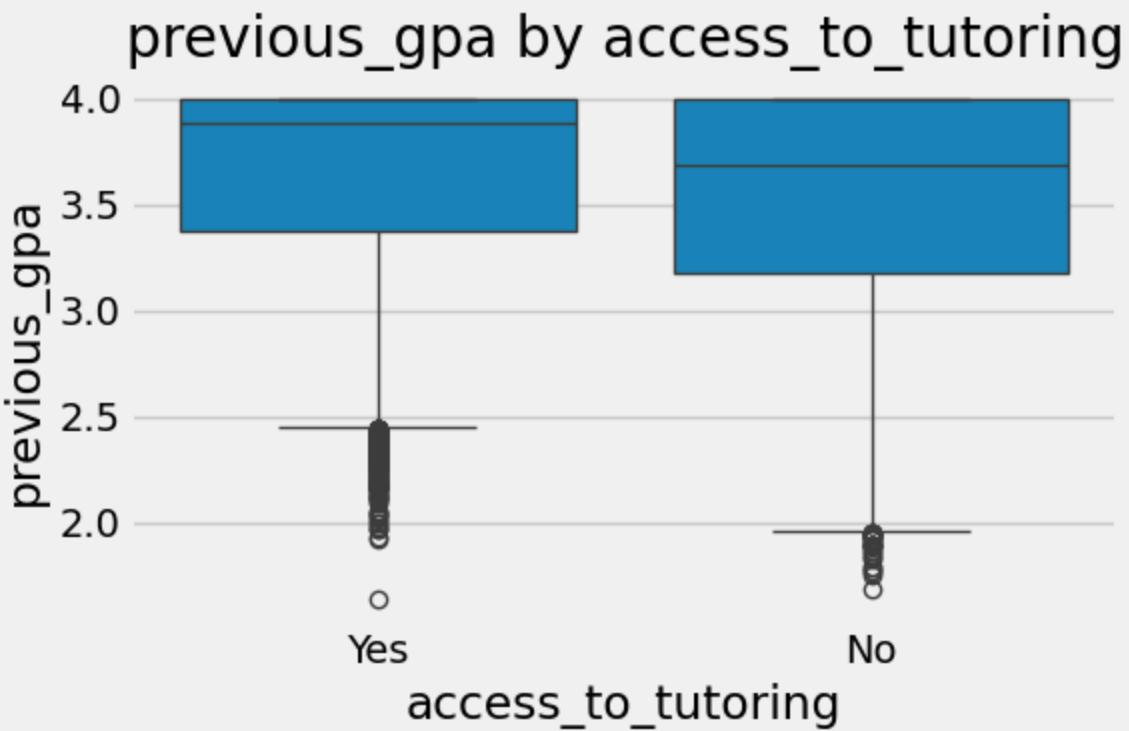


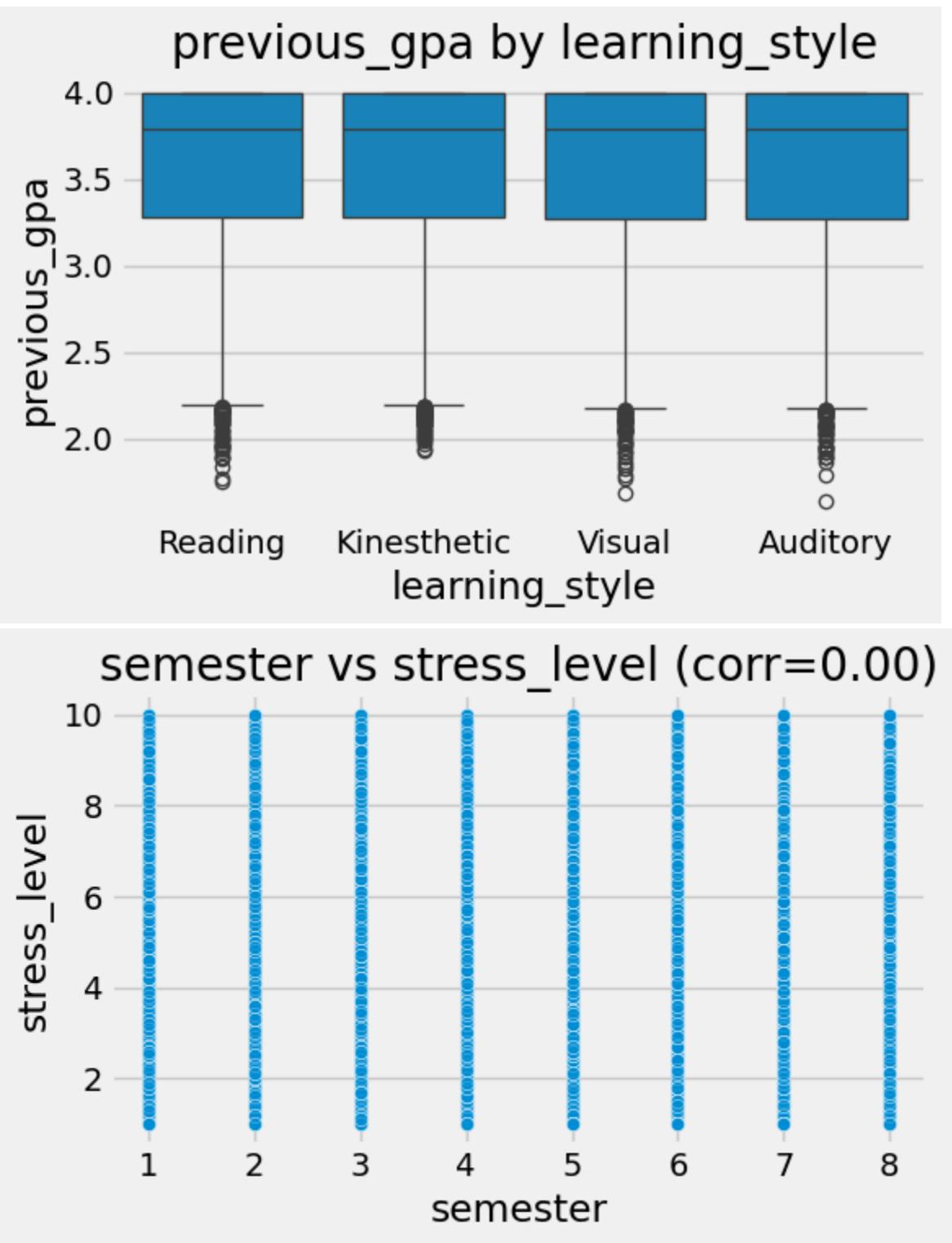




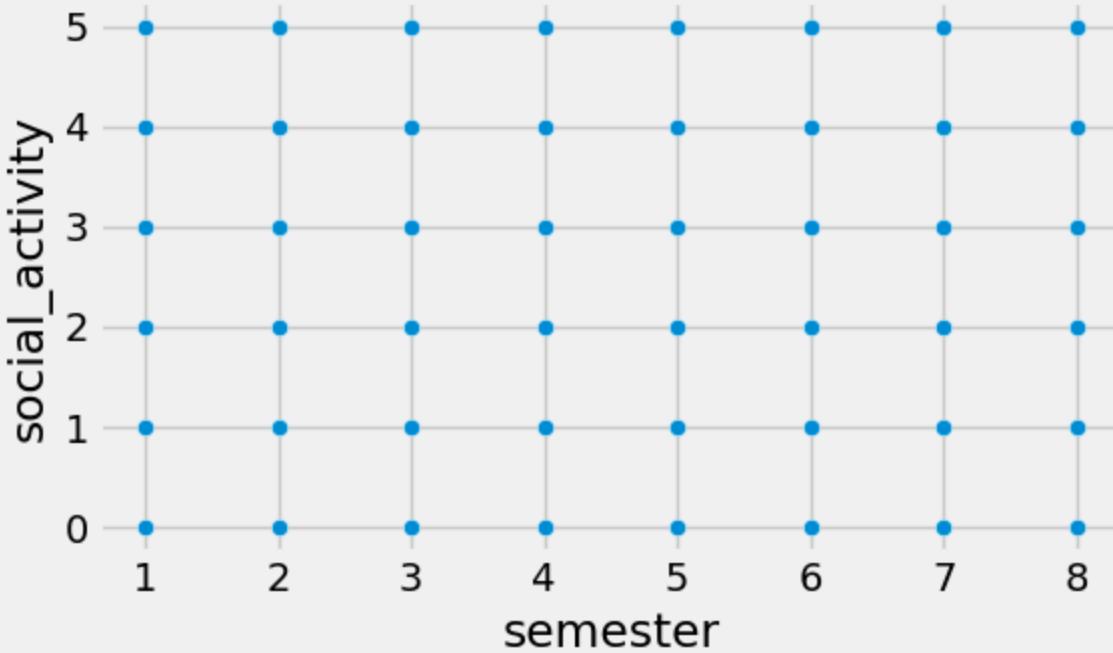




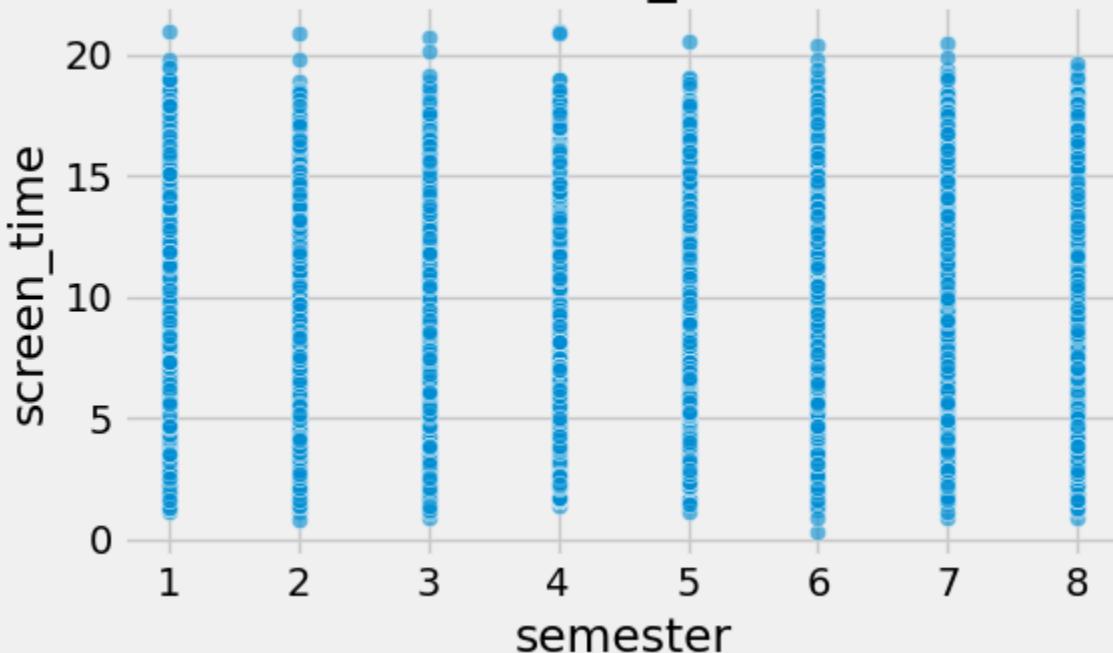




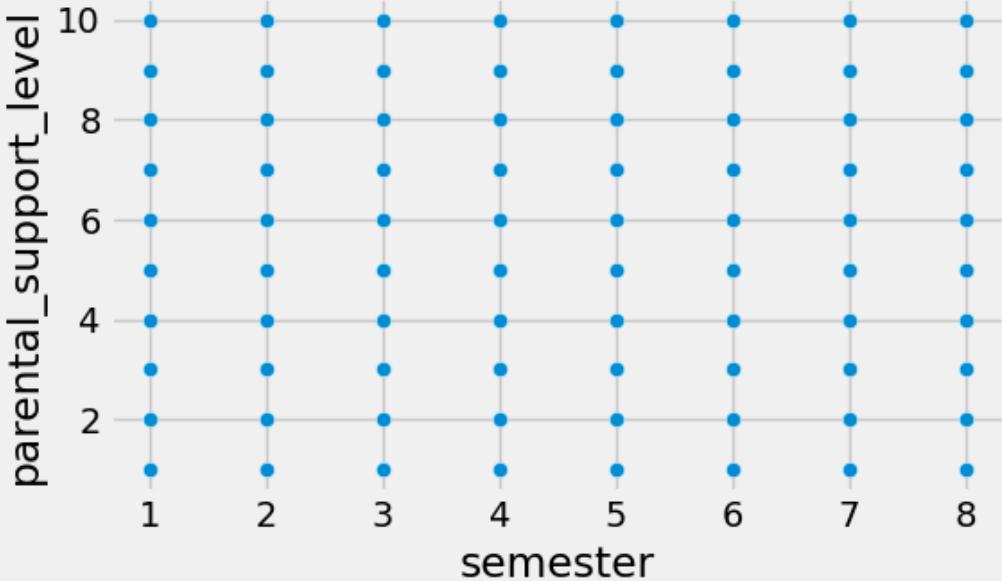
semester vs social_activity (corr=0.00)



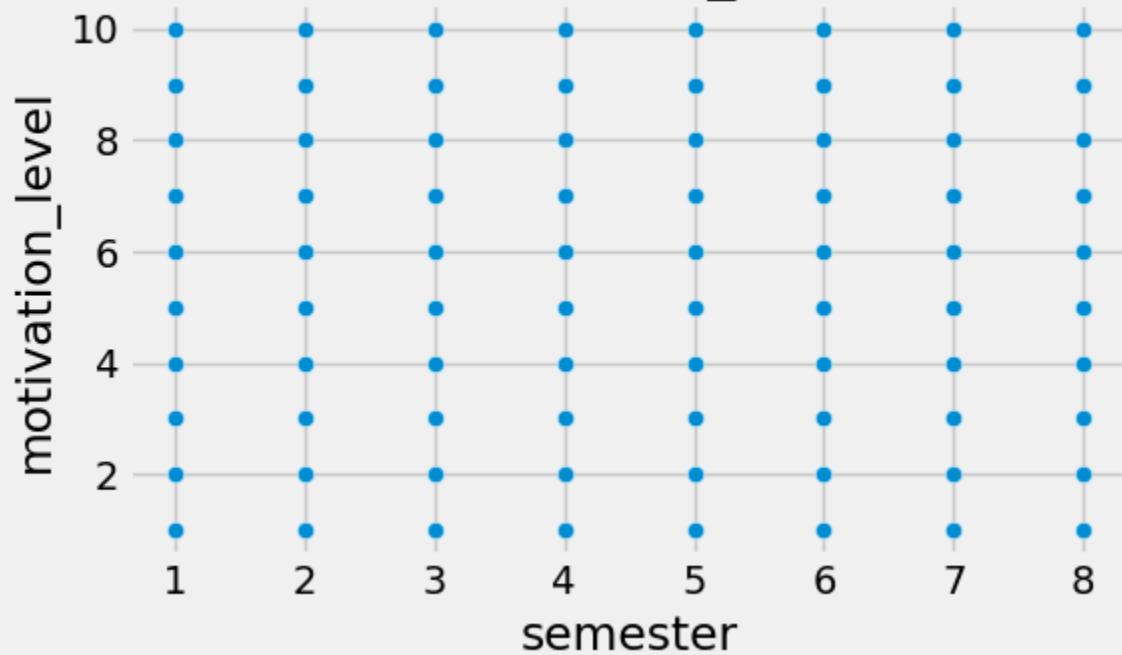
semester vs screen_time (corr=0.00)



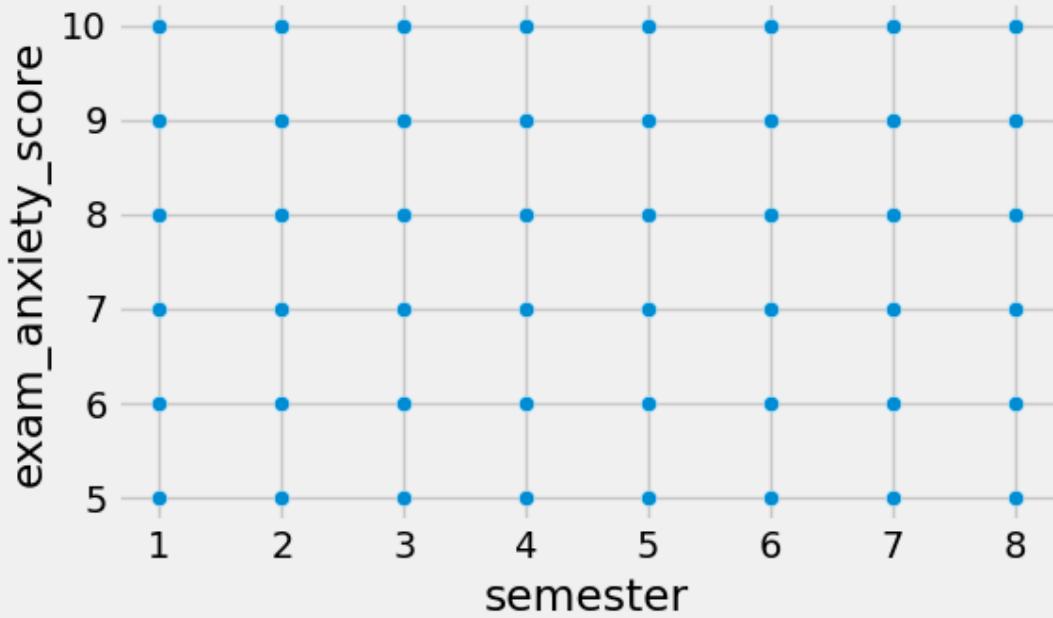
semester vs parental_support_level (corr=-0.00)



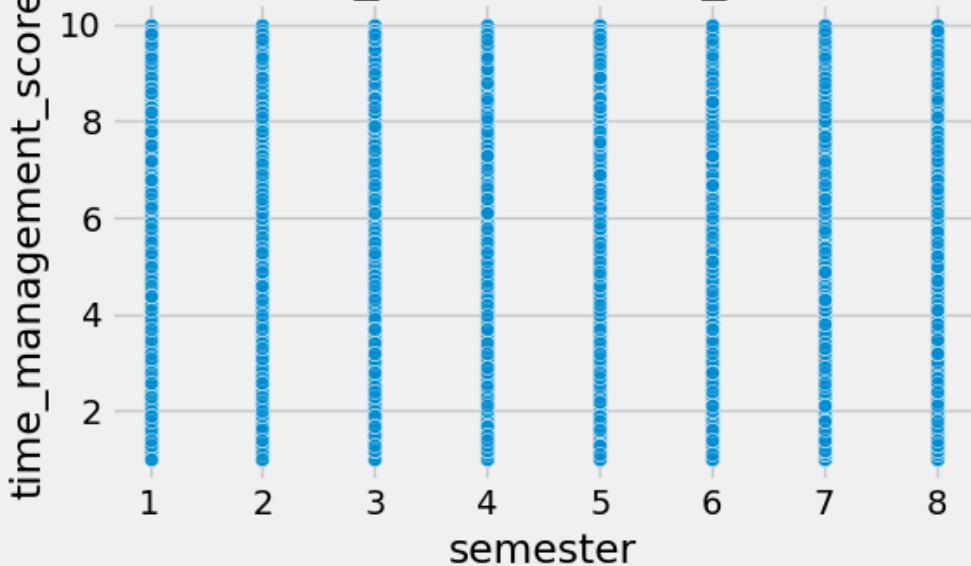
semester vs motivation_level (corr=0.00)

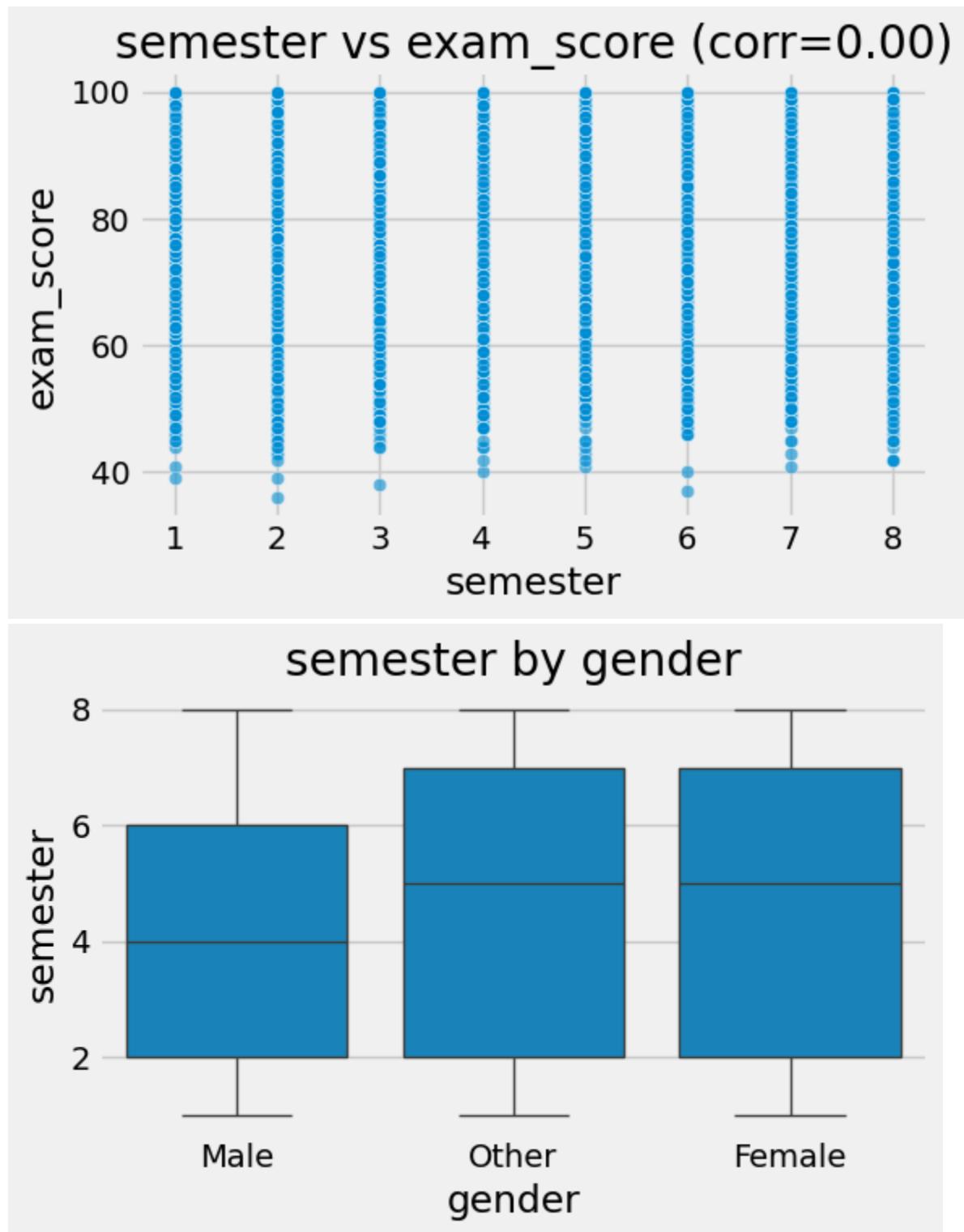


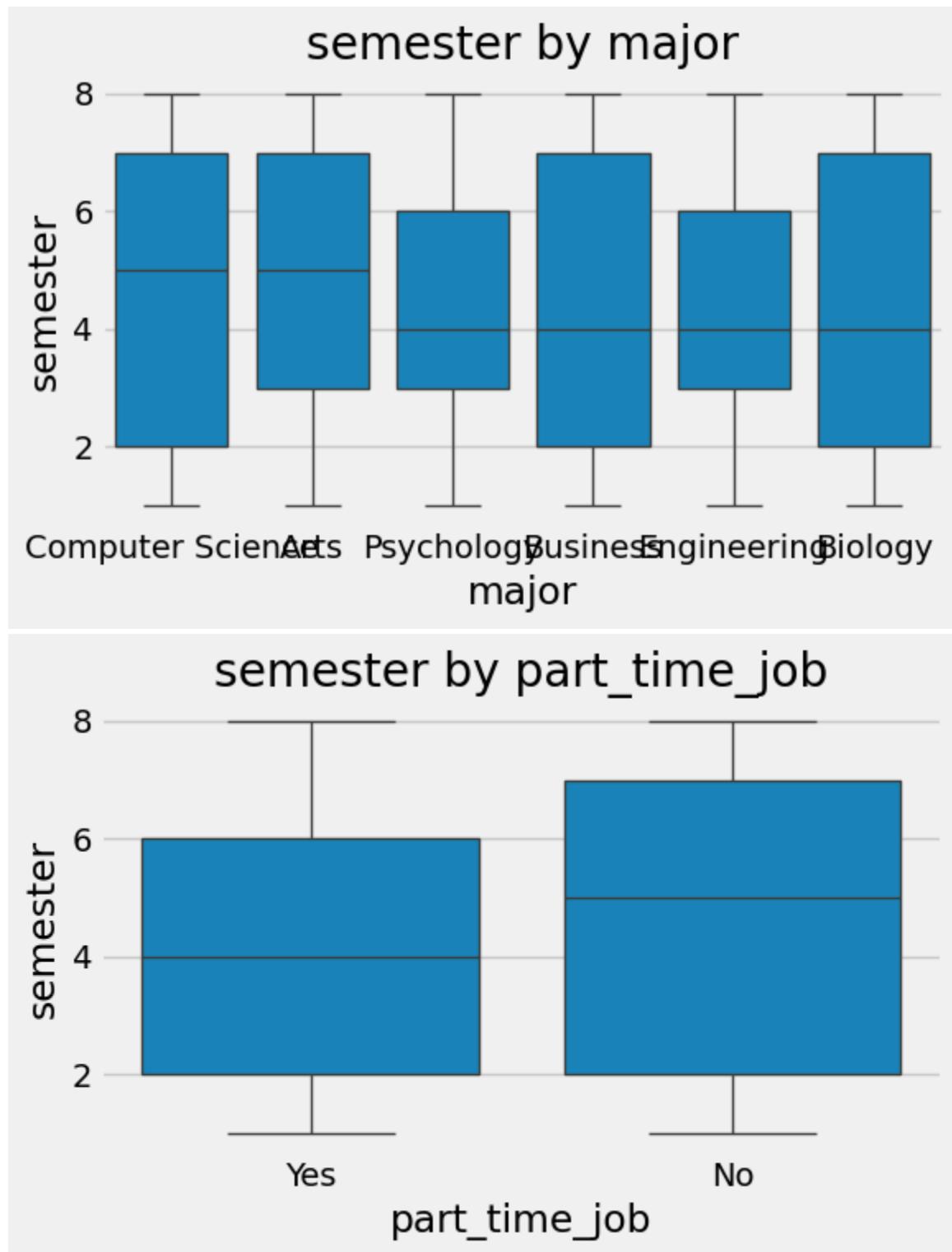
semester vs exam_anxiety_score (corr=-0.00)

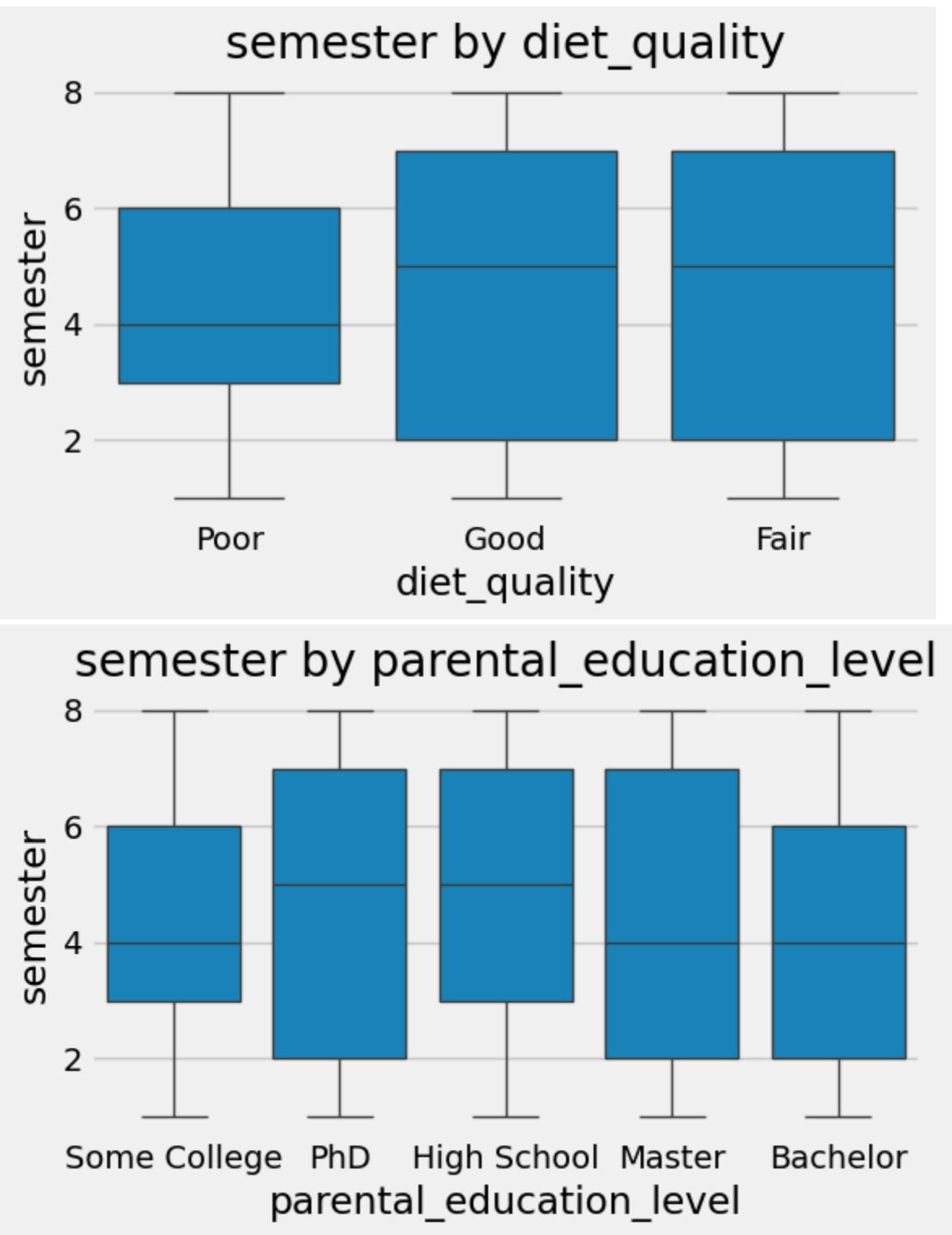


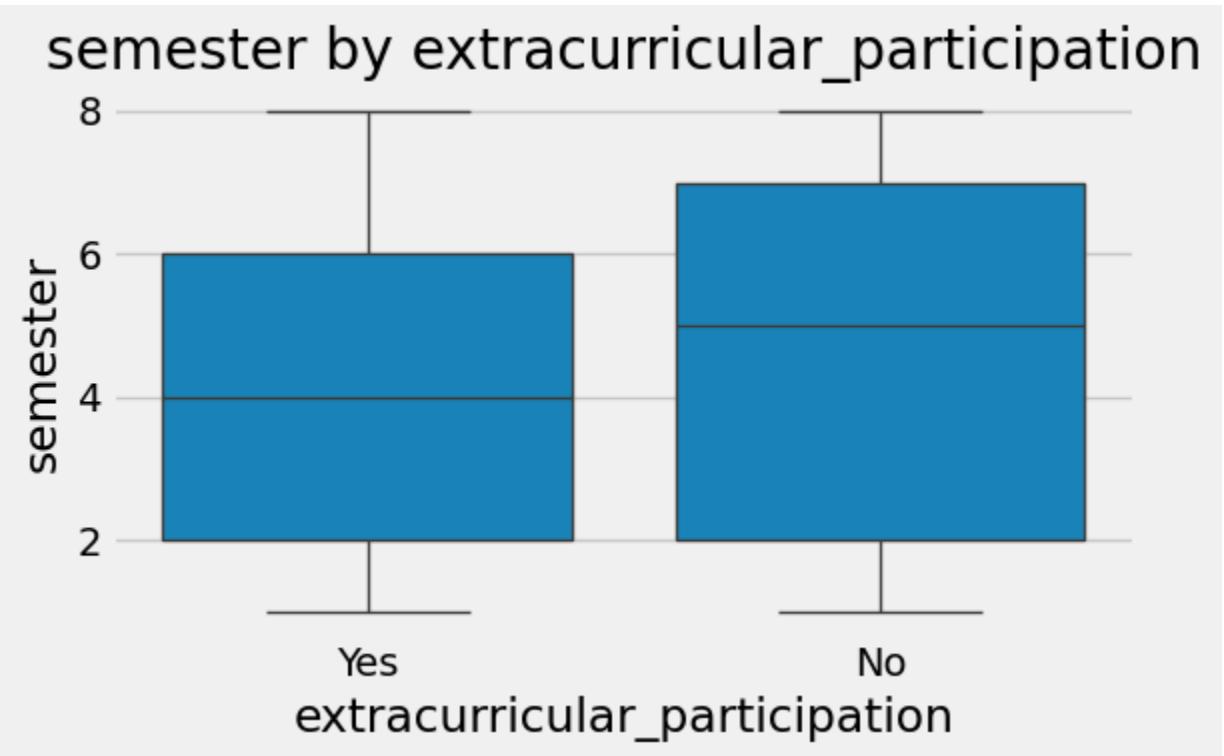
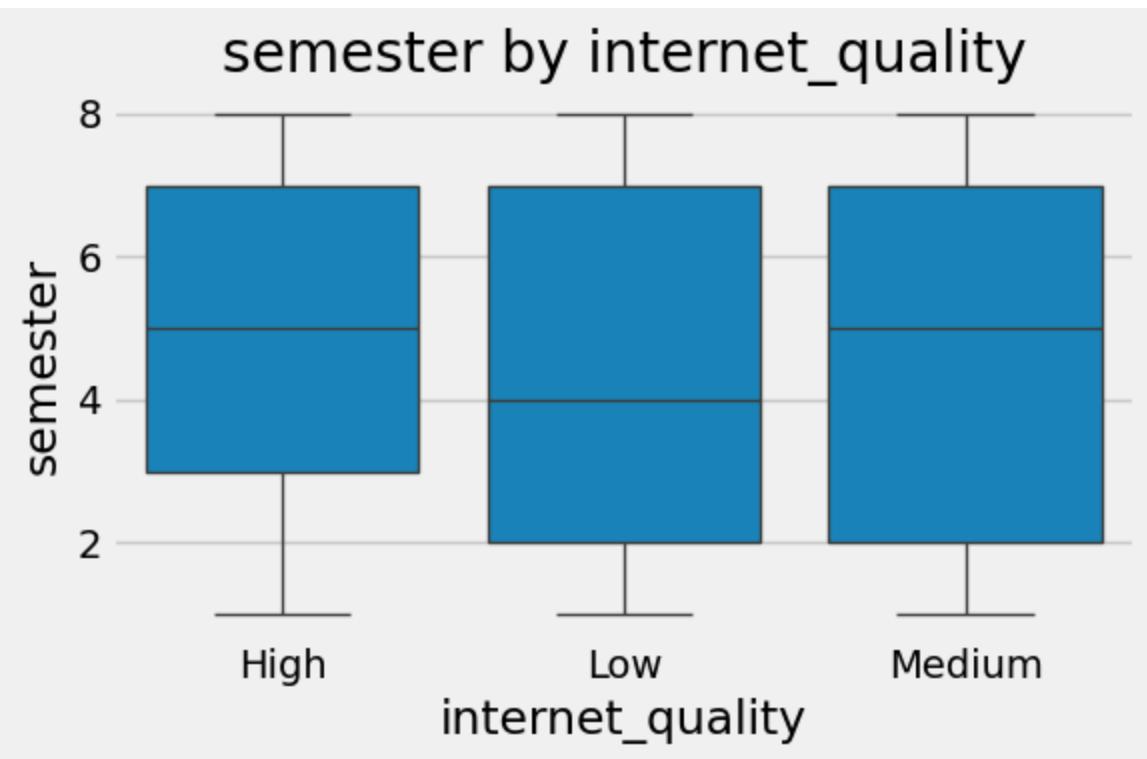
semester vs time_management_score (corr=0.00)

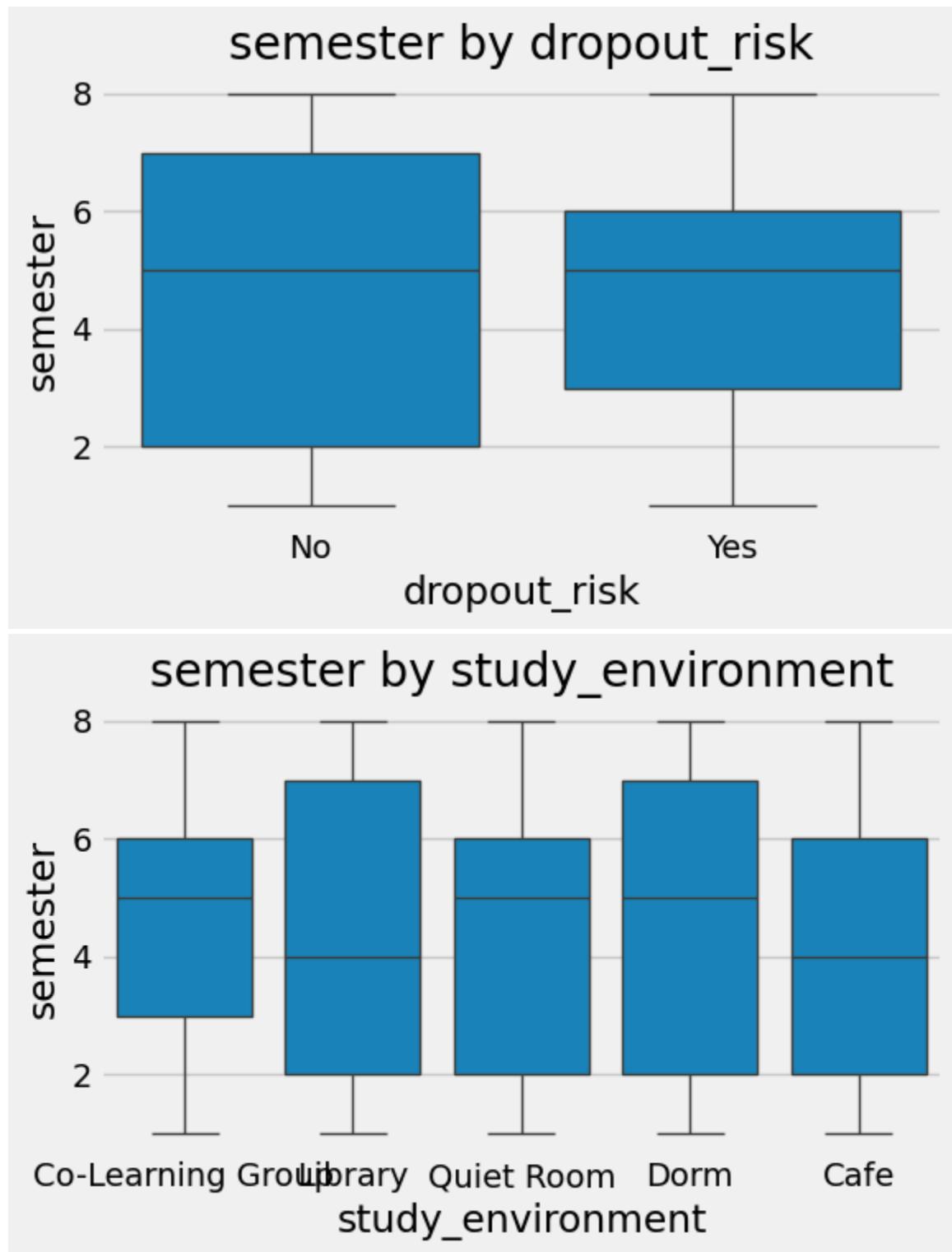


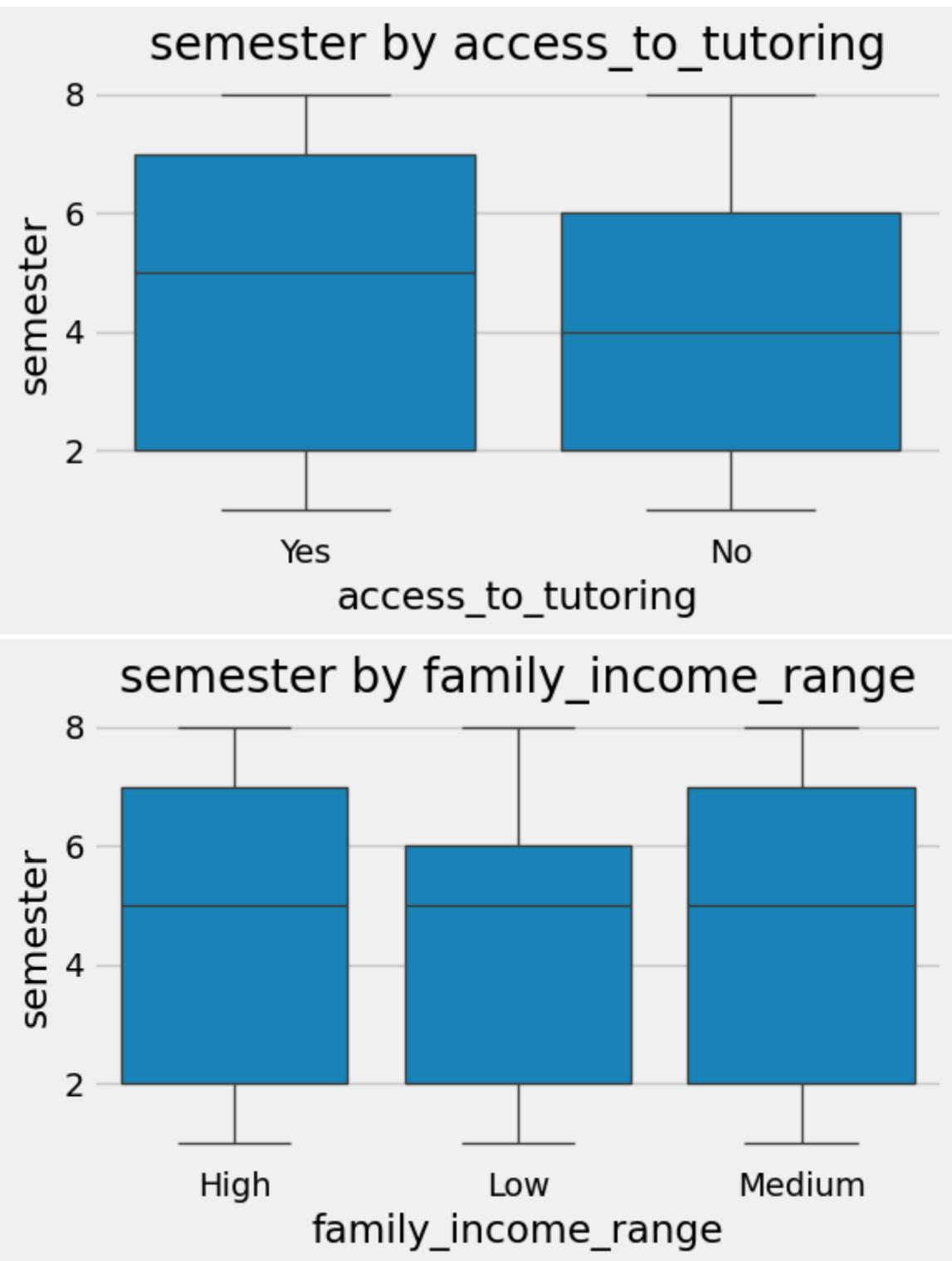


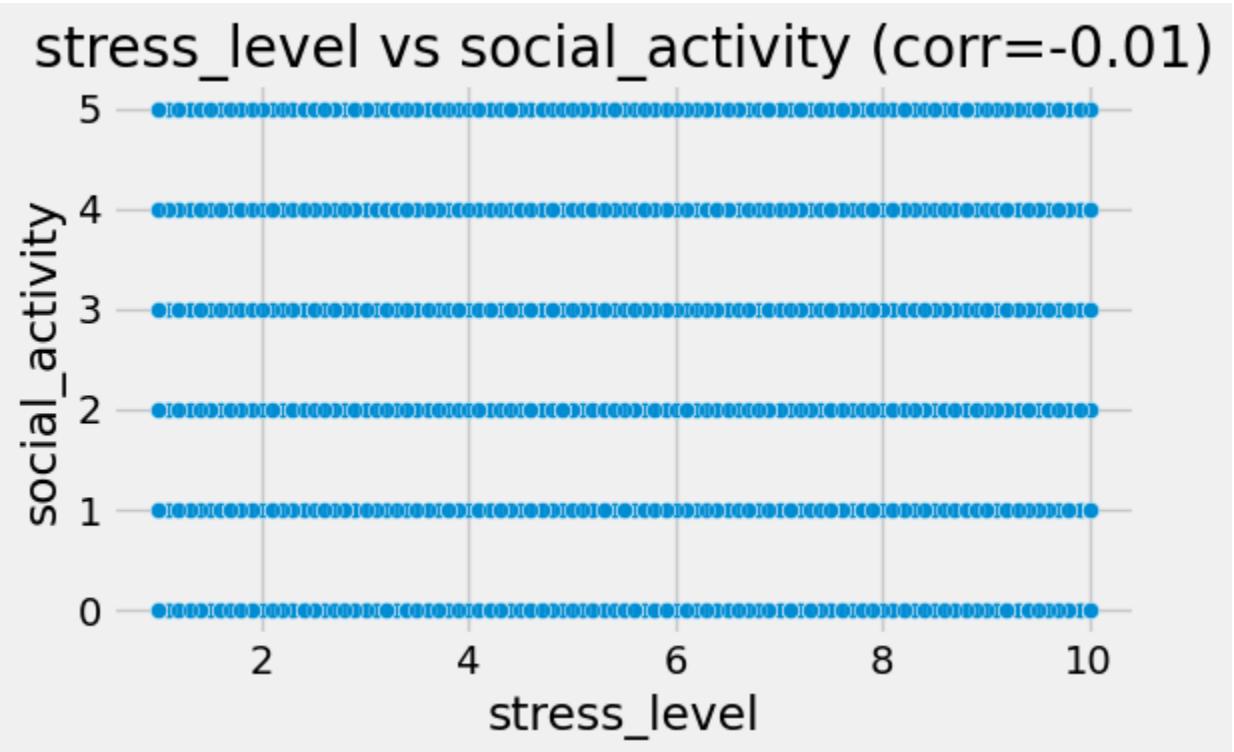
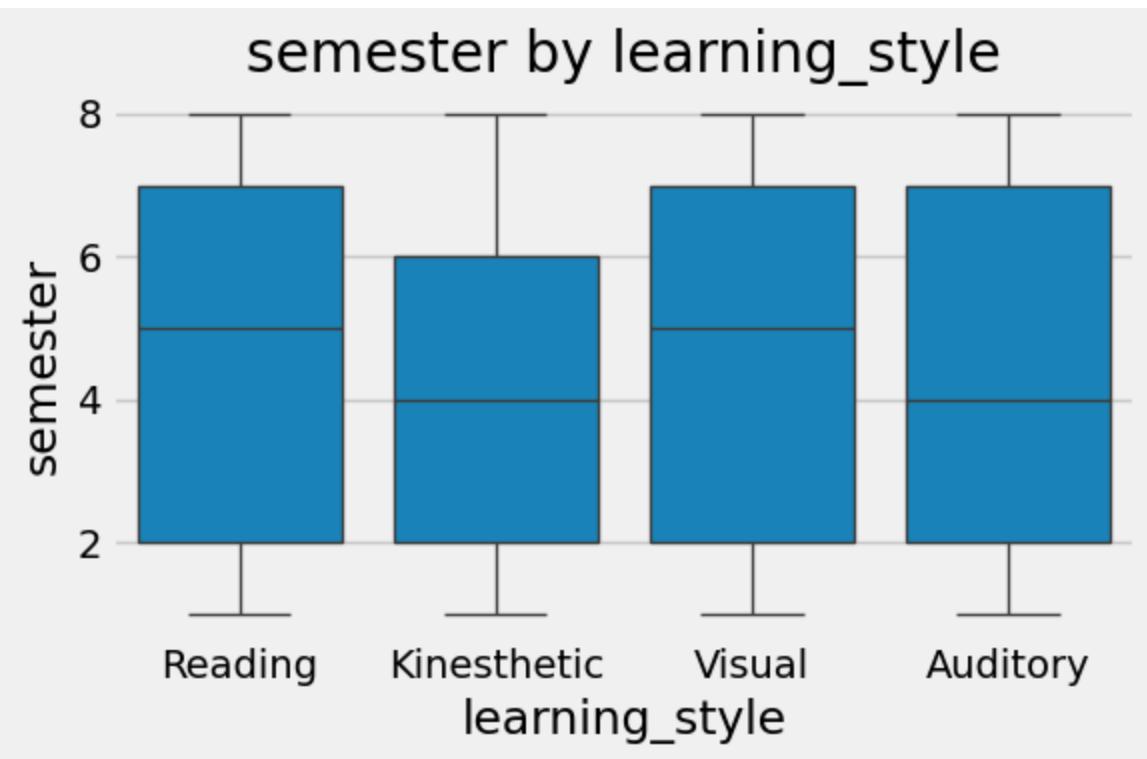




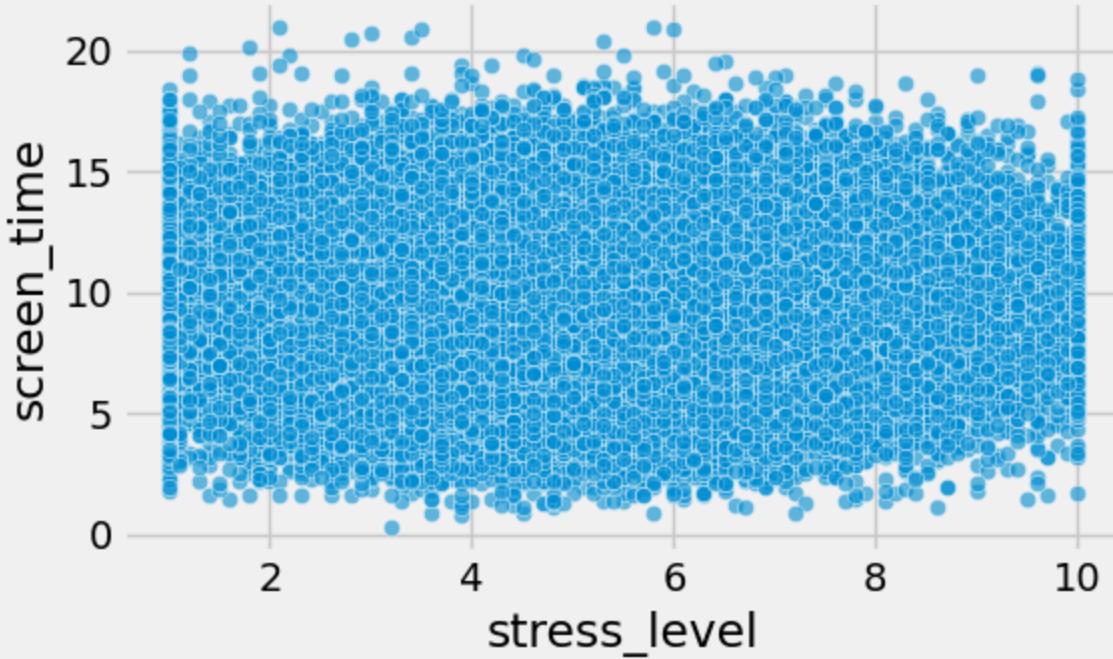




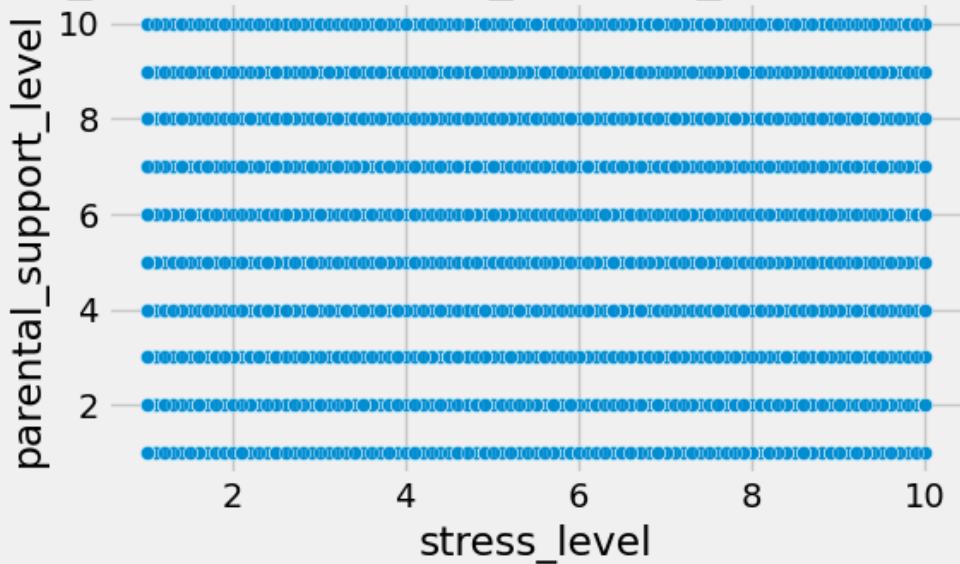




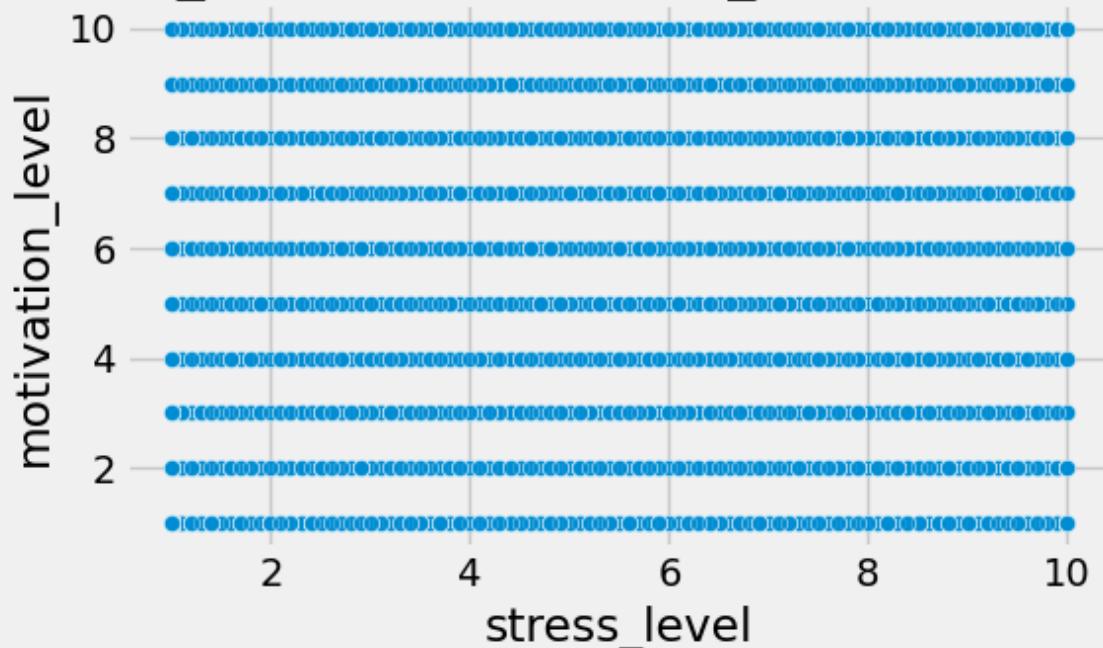
stress_level vs screen_time (corr=0.00)



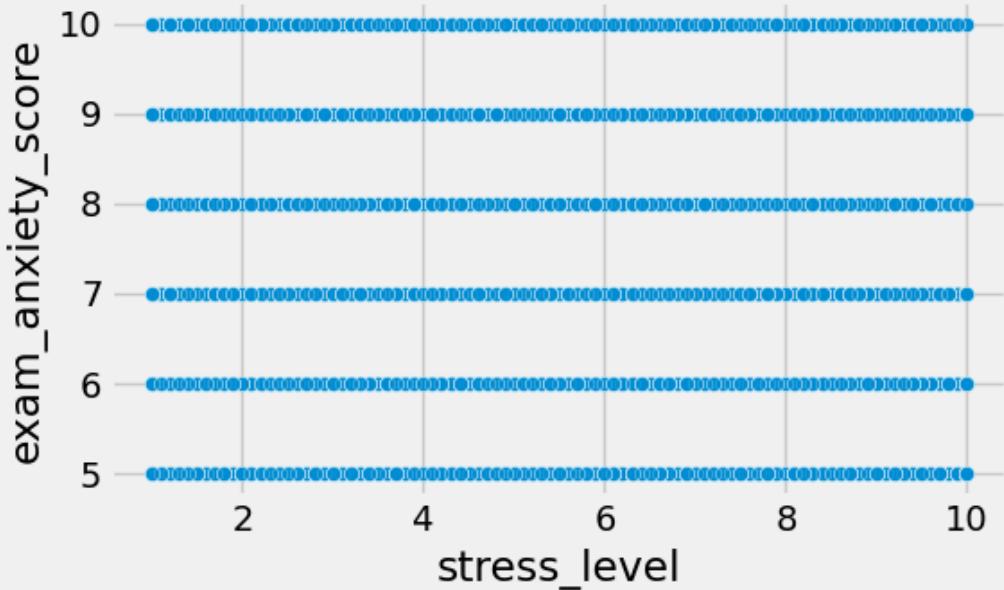
stress_level vs parental_support_level (corr=-0.00)



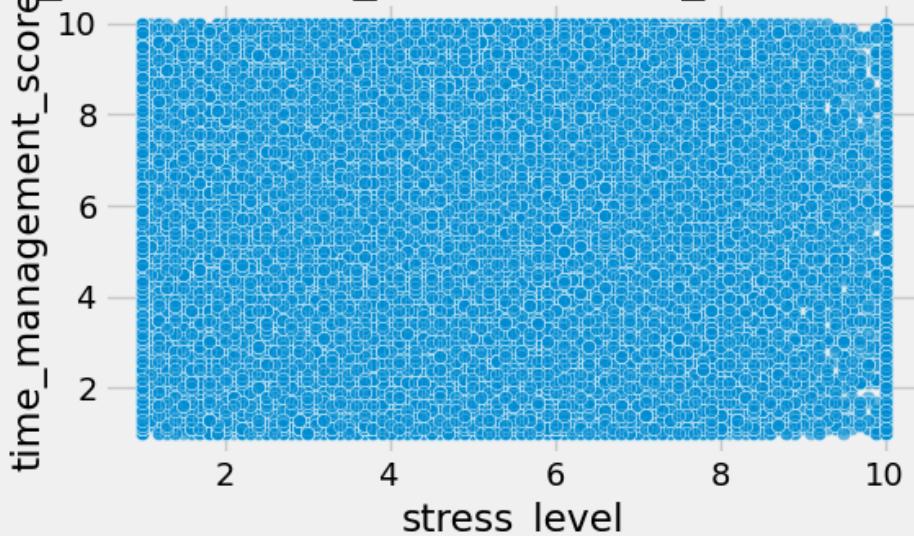
stress_level vs motivation_level (corr=0.00)



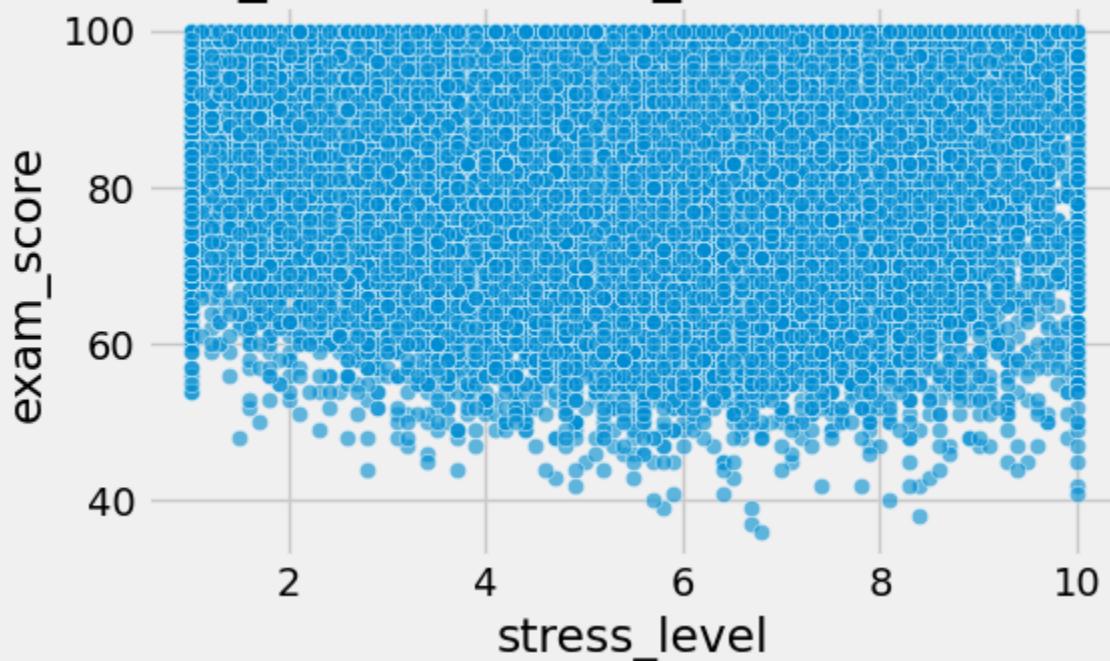
stress_level vs exam_anxiety_score (corr=-0.00)



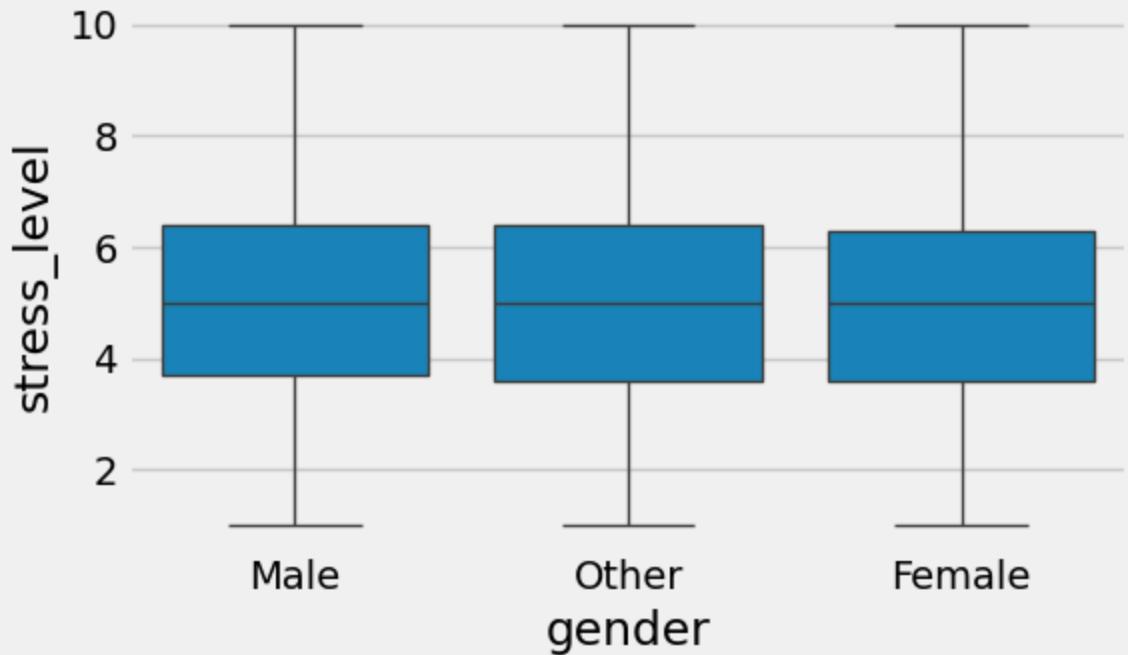
stress_level vs time_management_score (corr=-0.00)



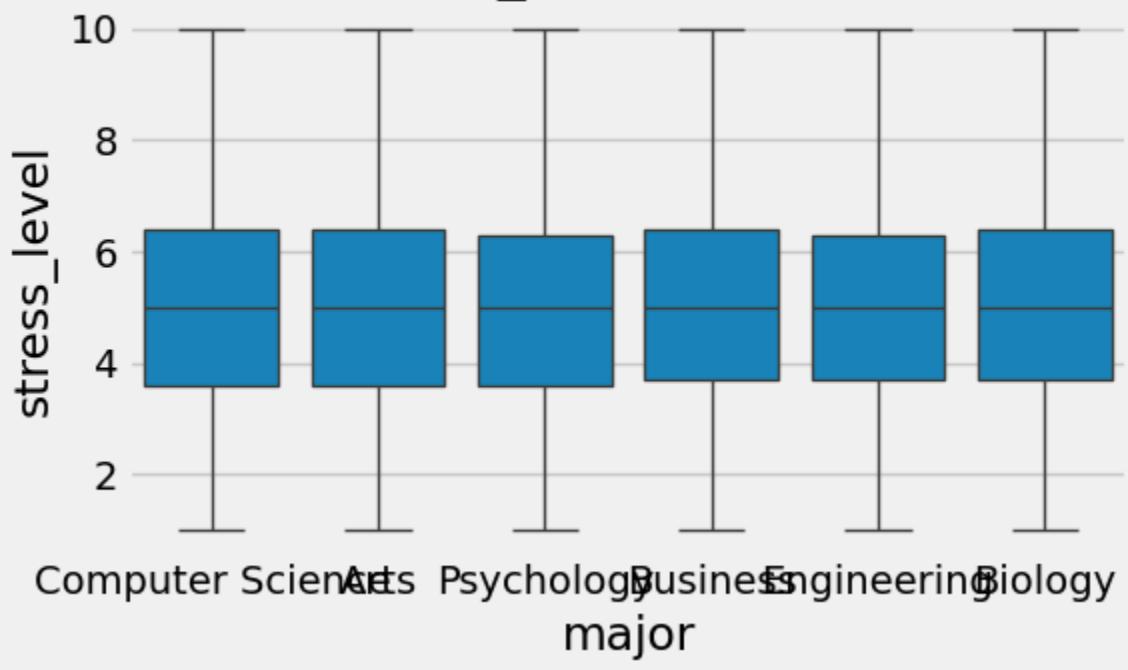
stress_level vs exam_score (corr=-0.12)

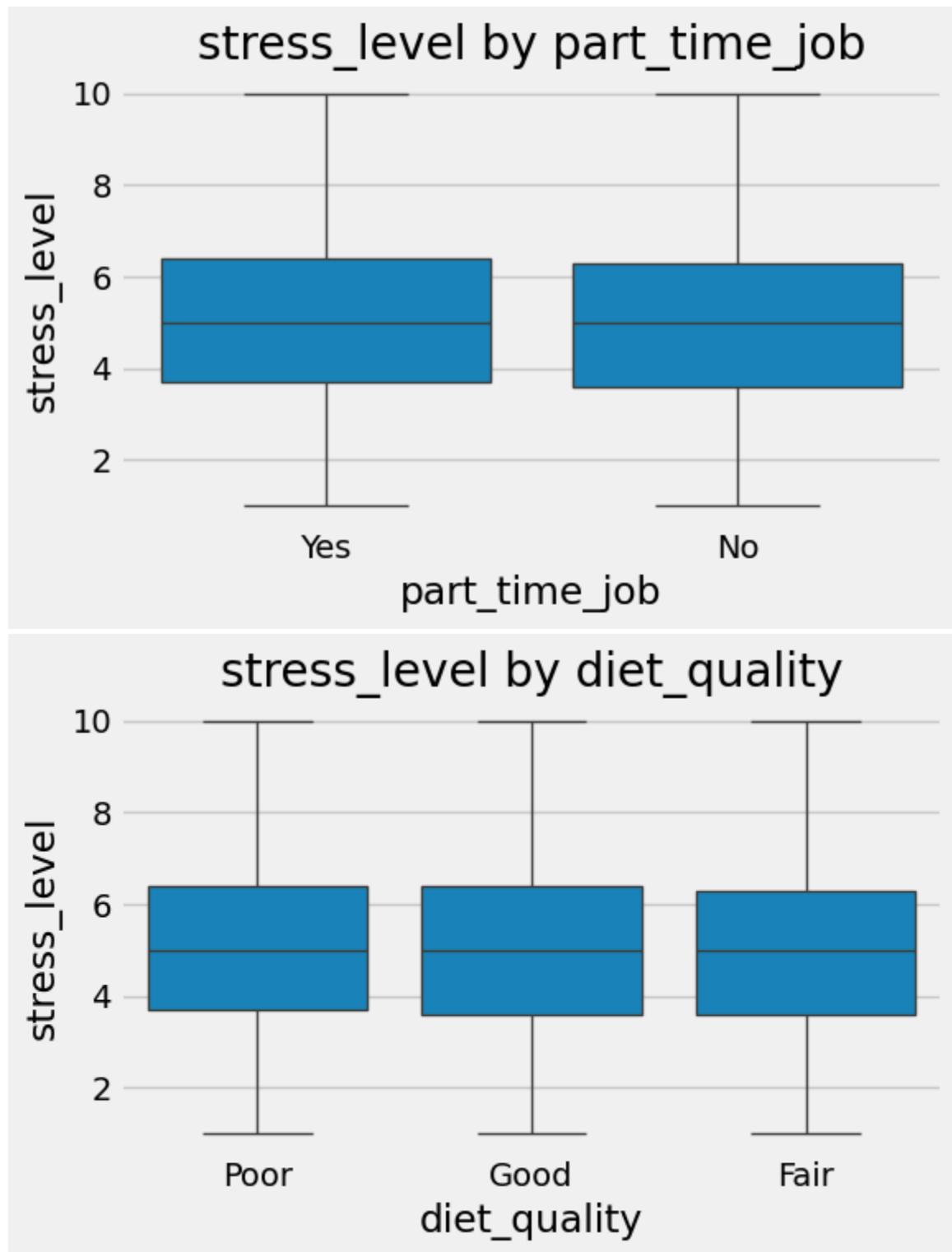


stress_level by gender

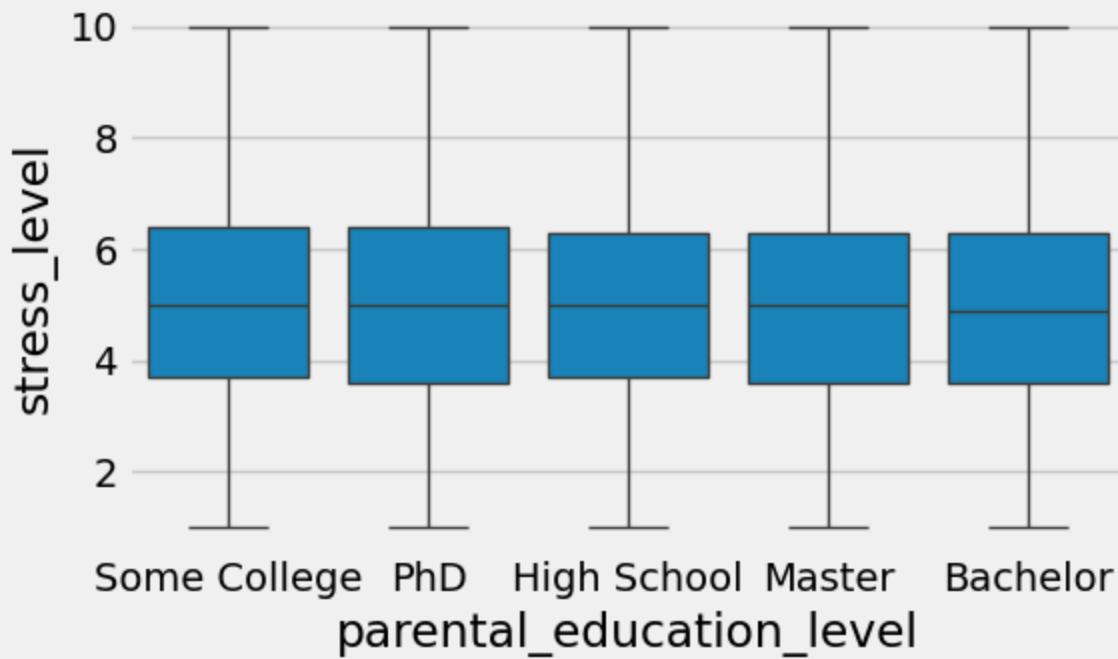


stress_level by major

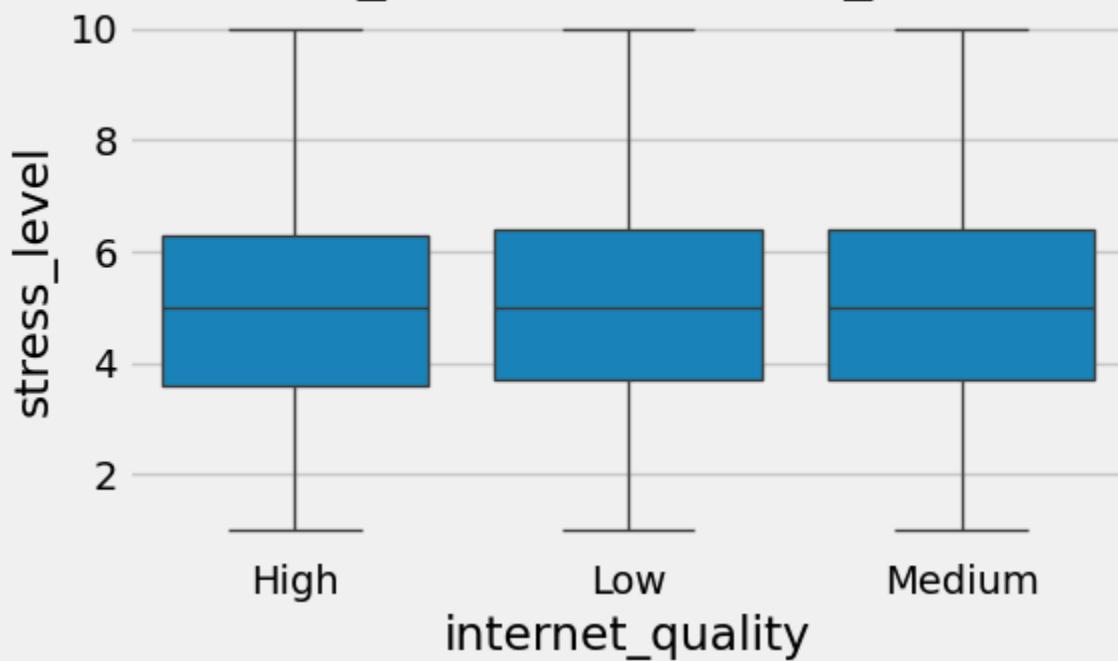


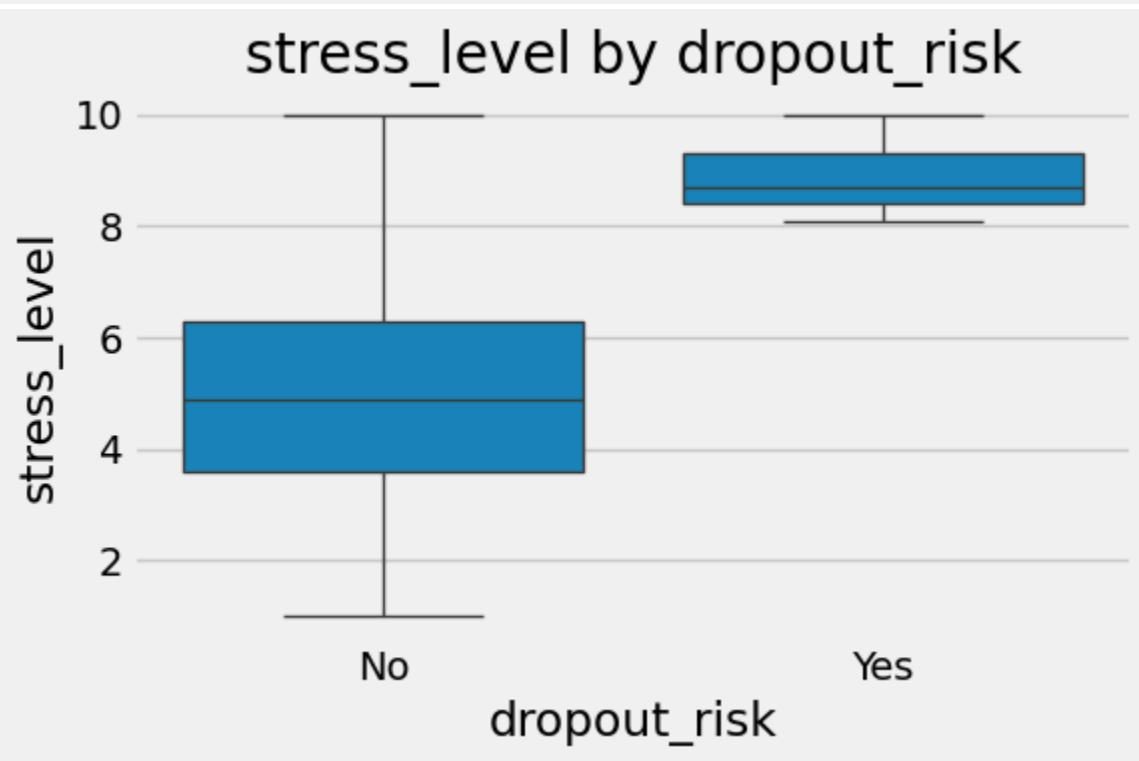
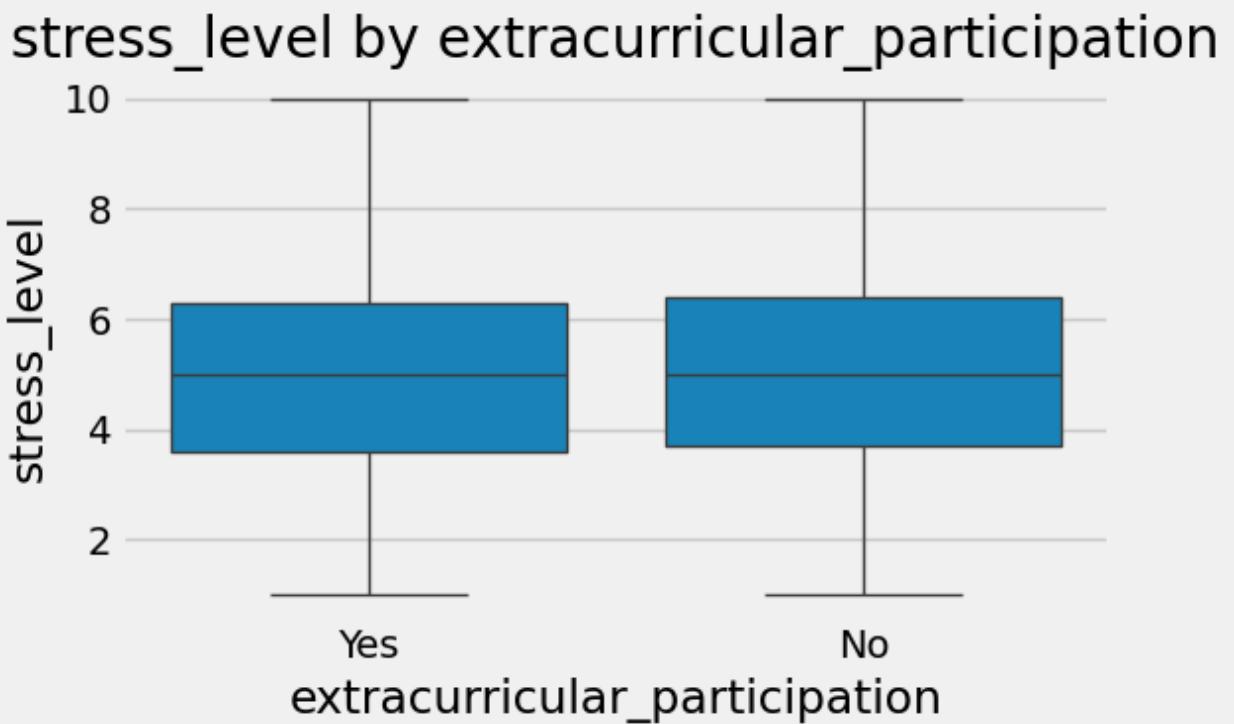


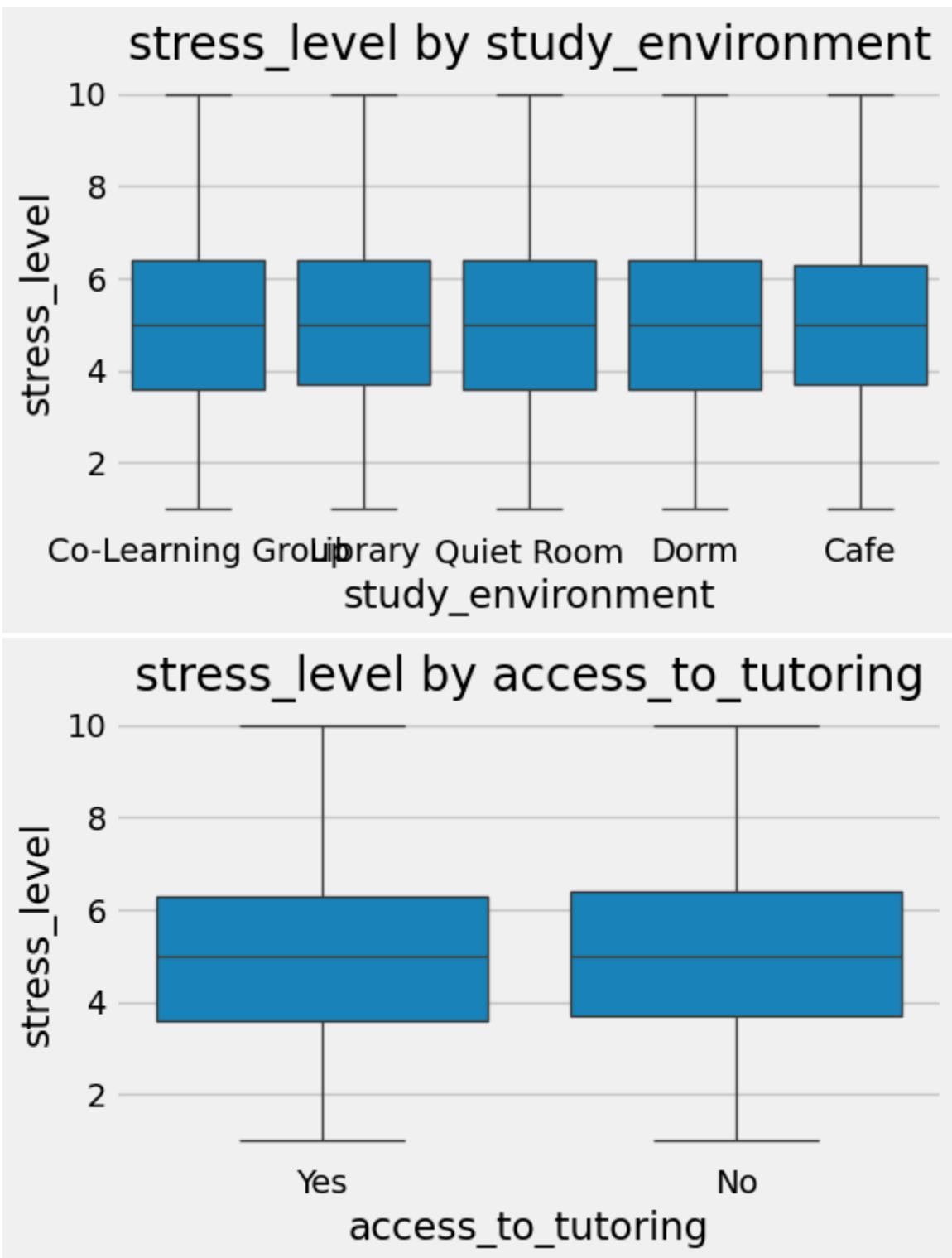
stress_level by parental_education_level

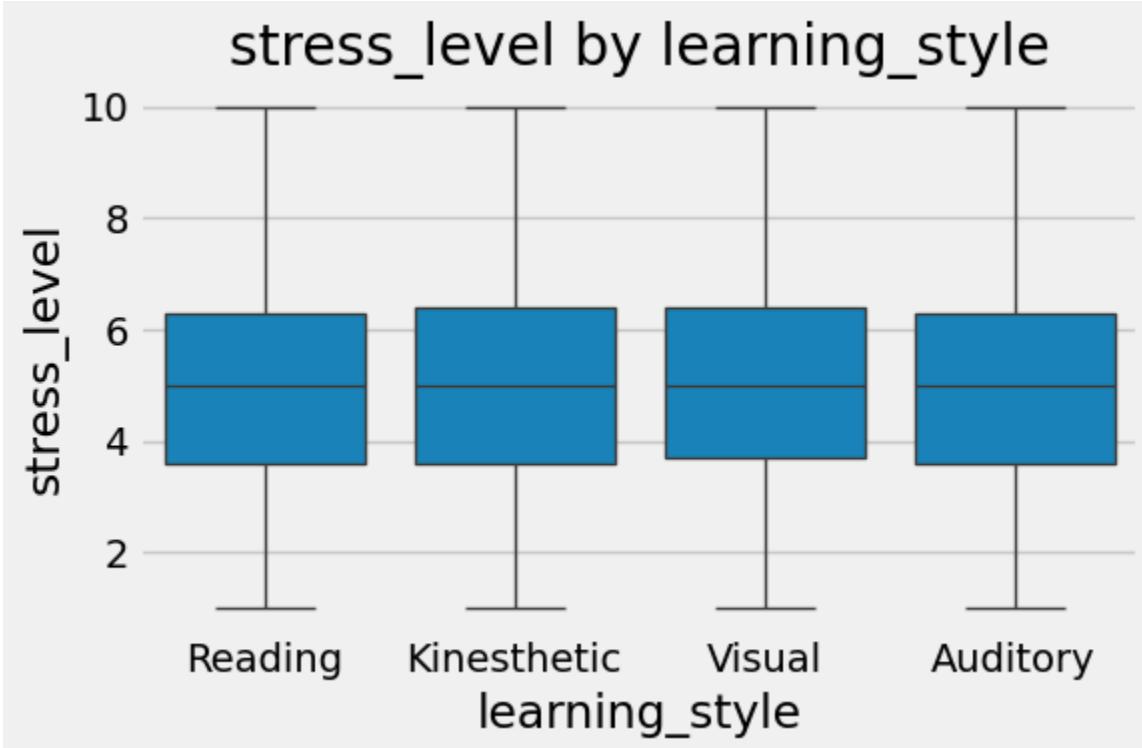
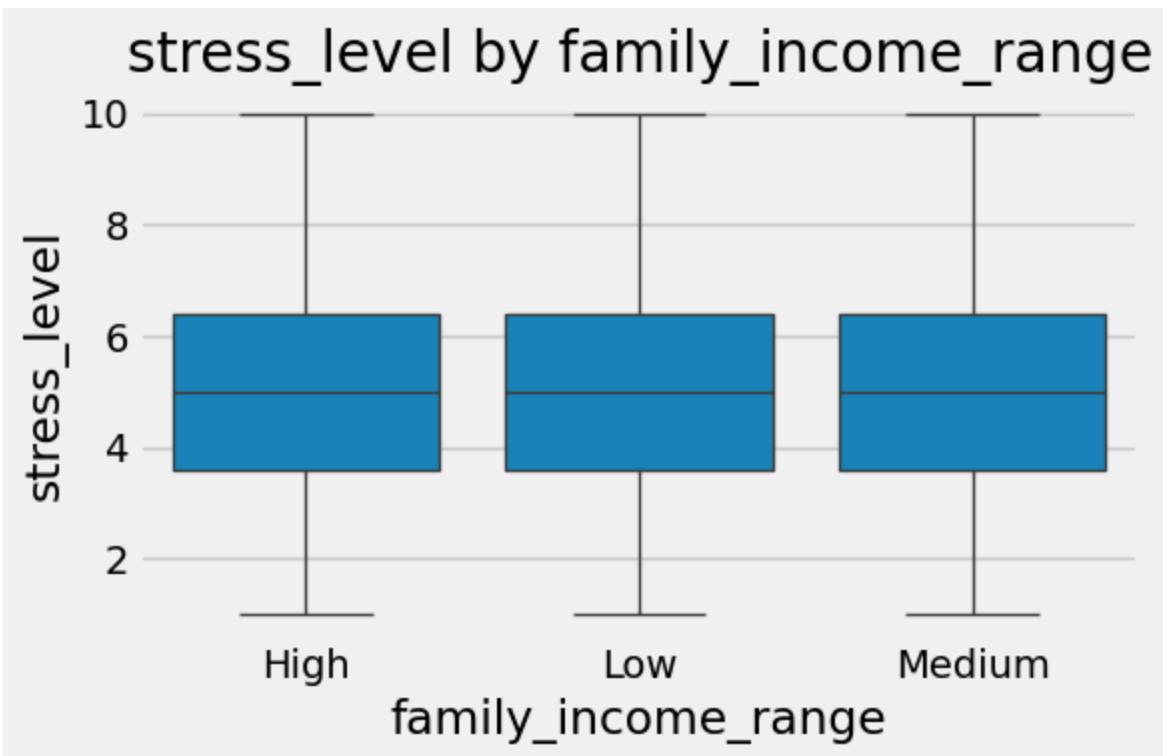


stress_level by internet_quality

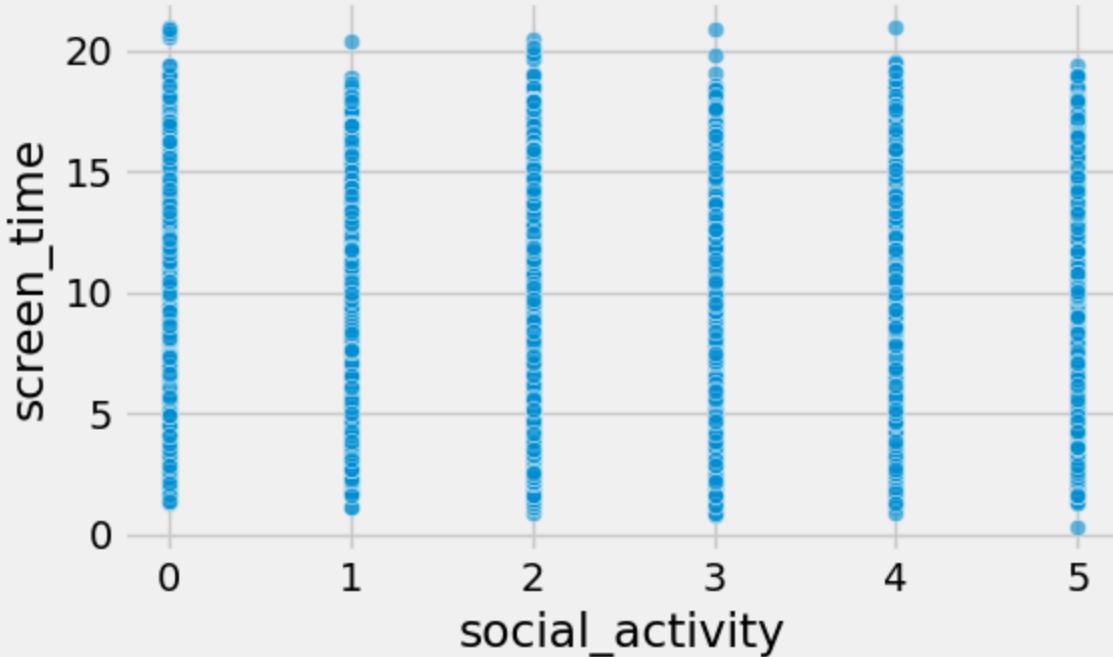




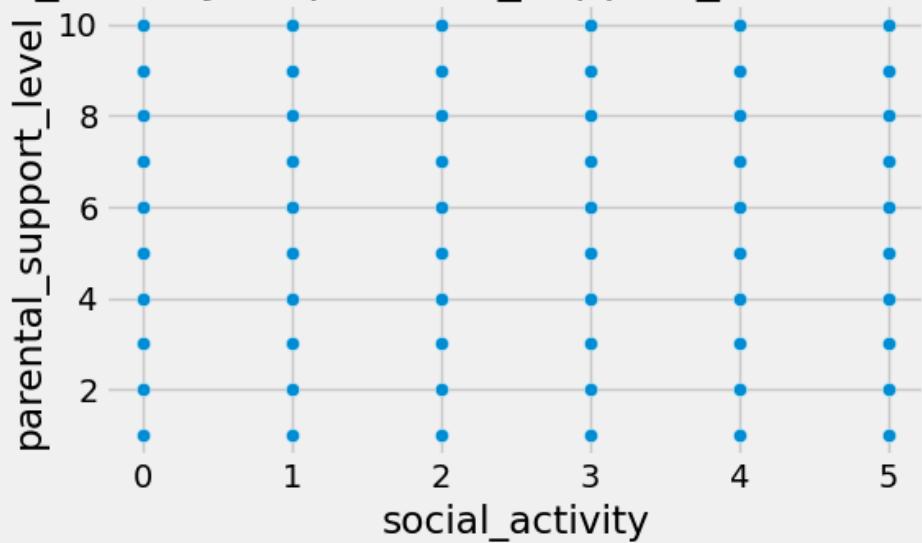




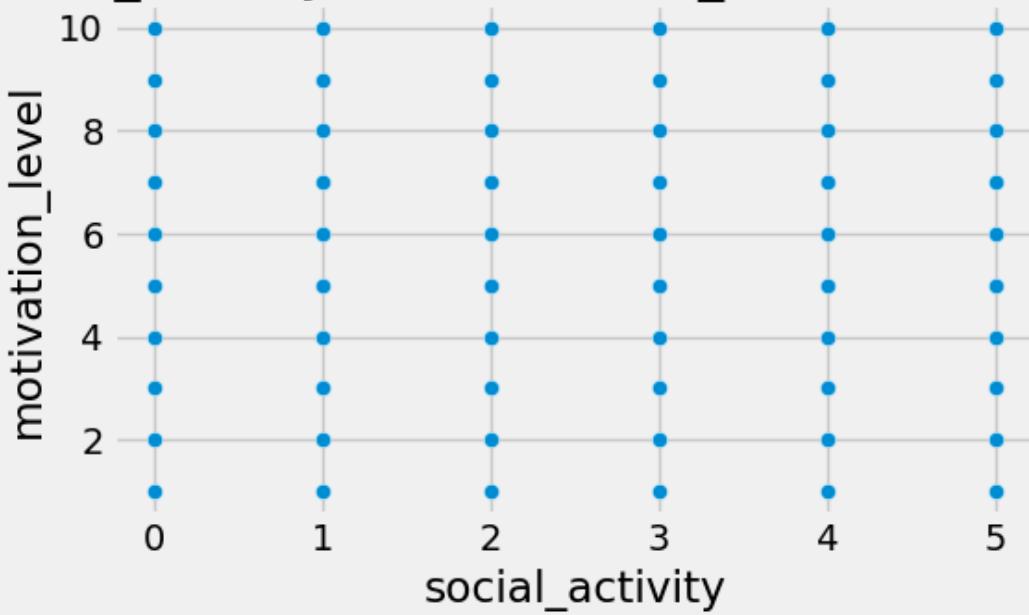
social_activity vs screen_time (corr=-0.00)



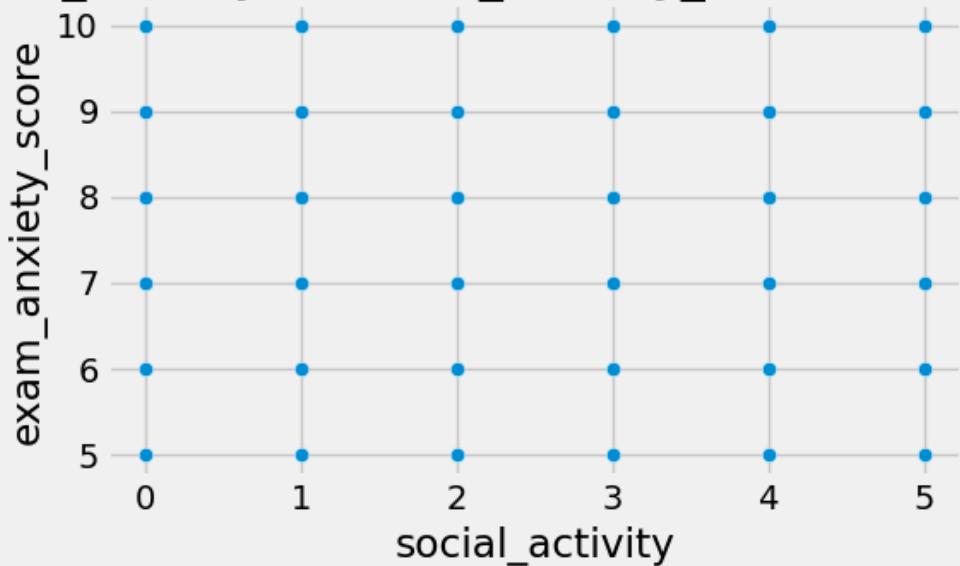
social_activity vs parental_support_level (corr=-0.00)



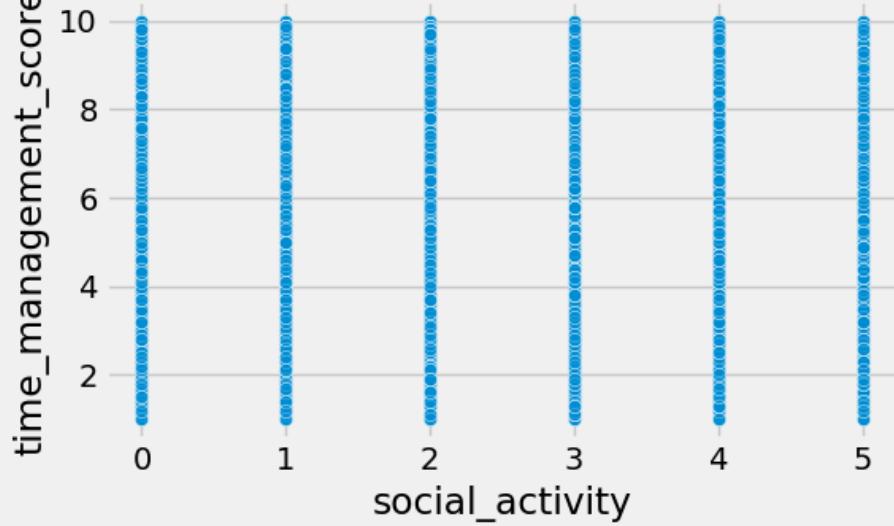
social_activity vs motivation_level (corr=-0.00)



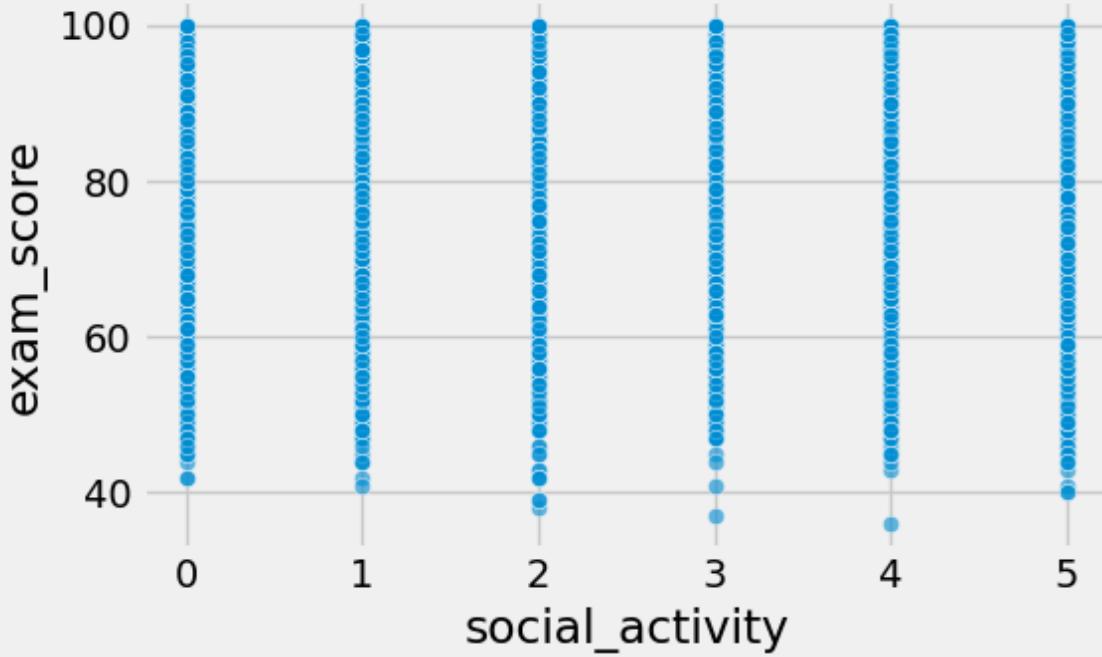
social_activity vs exam_anxiety_score (corr=-0.00)

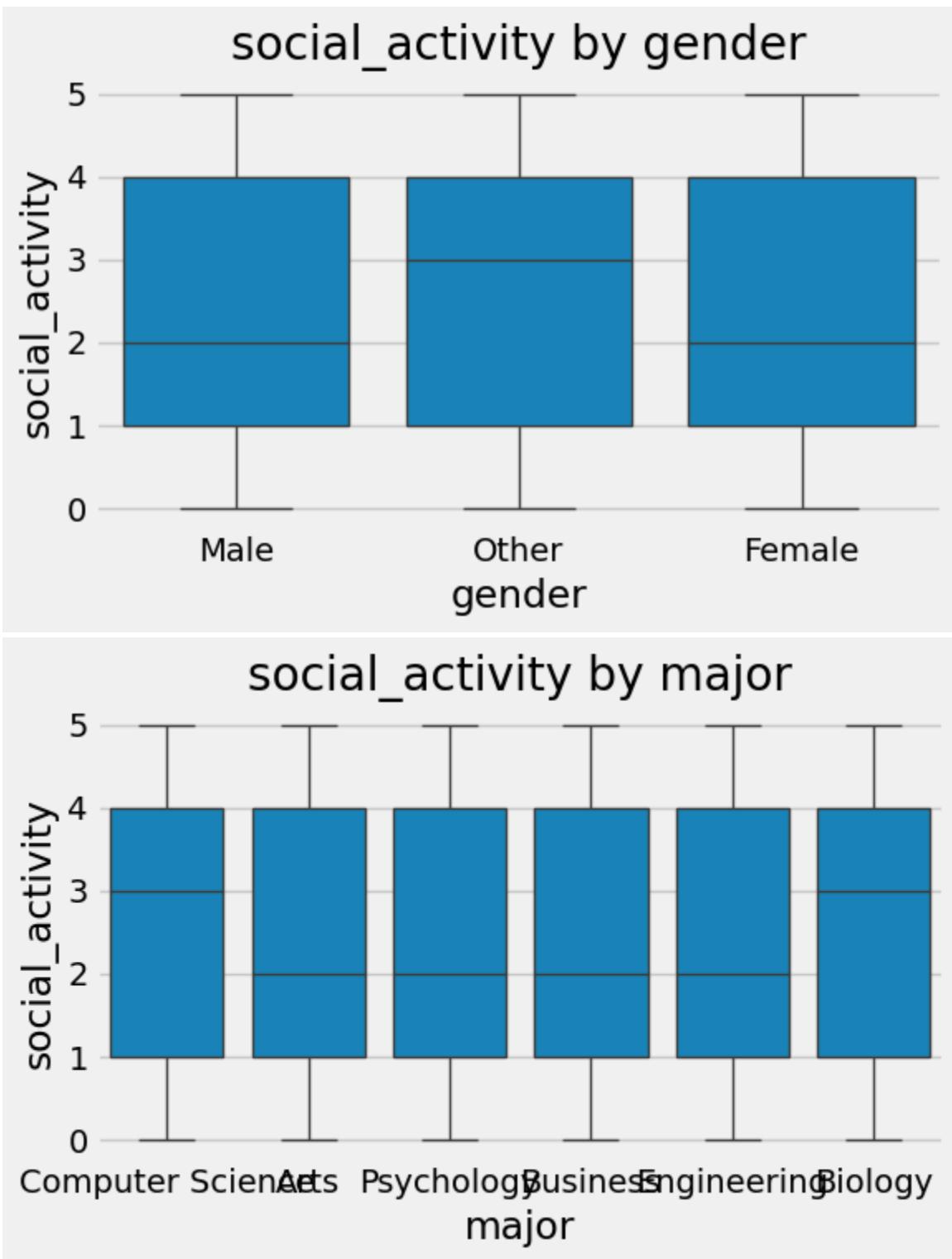


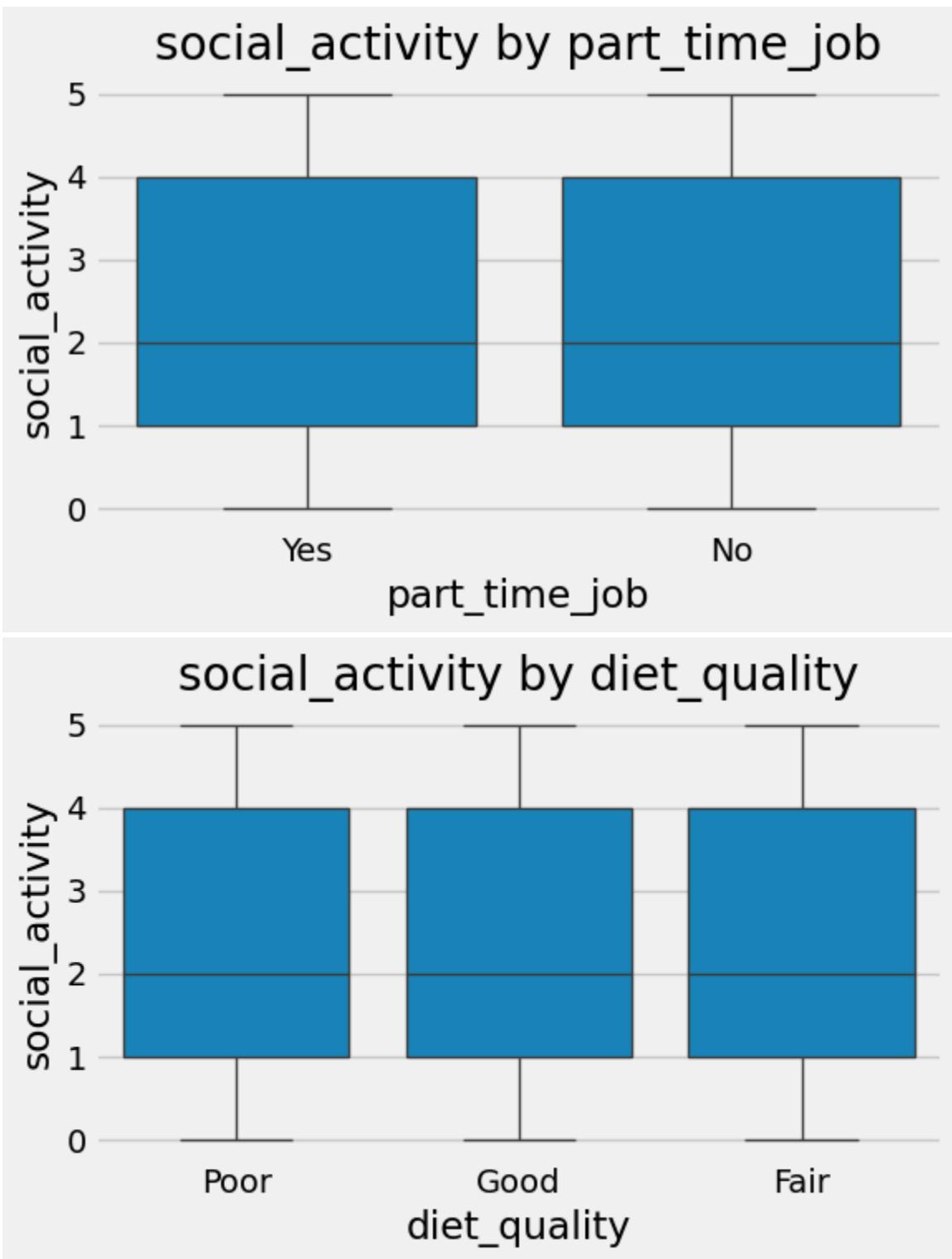
social_activity vs time_management_score (corr=0.00)



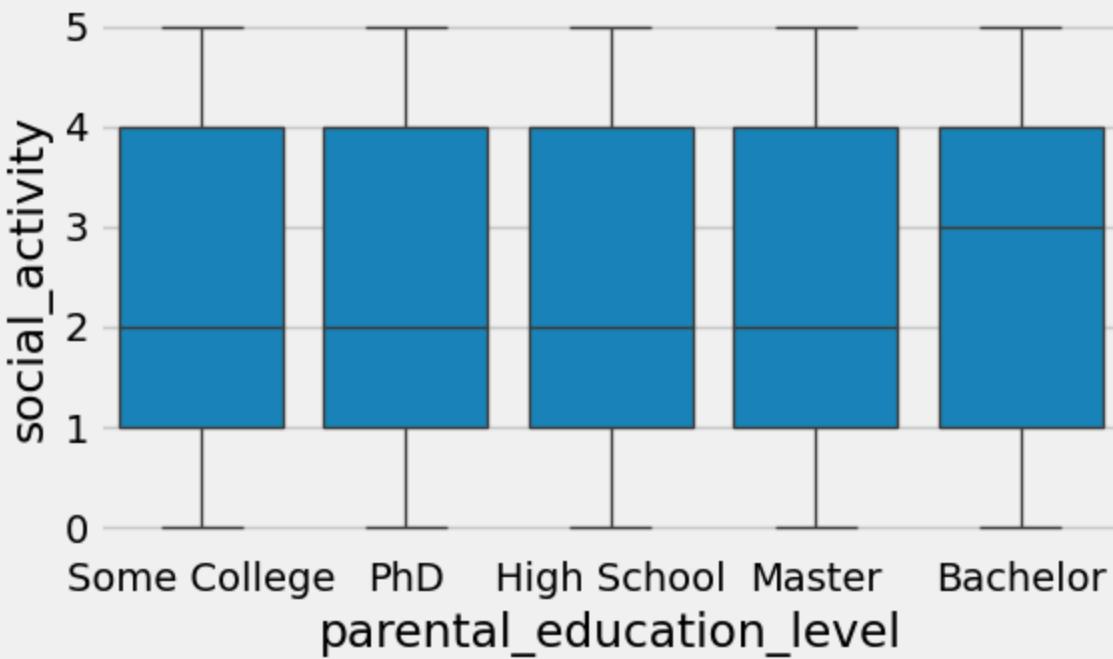
social_activity vs exam_score (corr=-0.00)



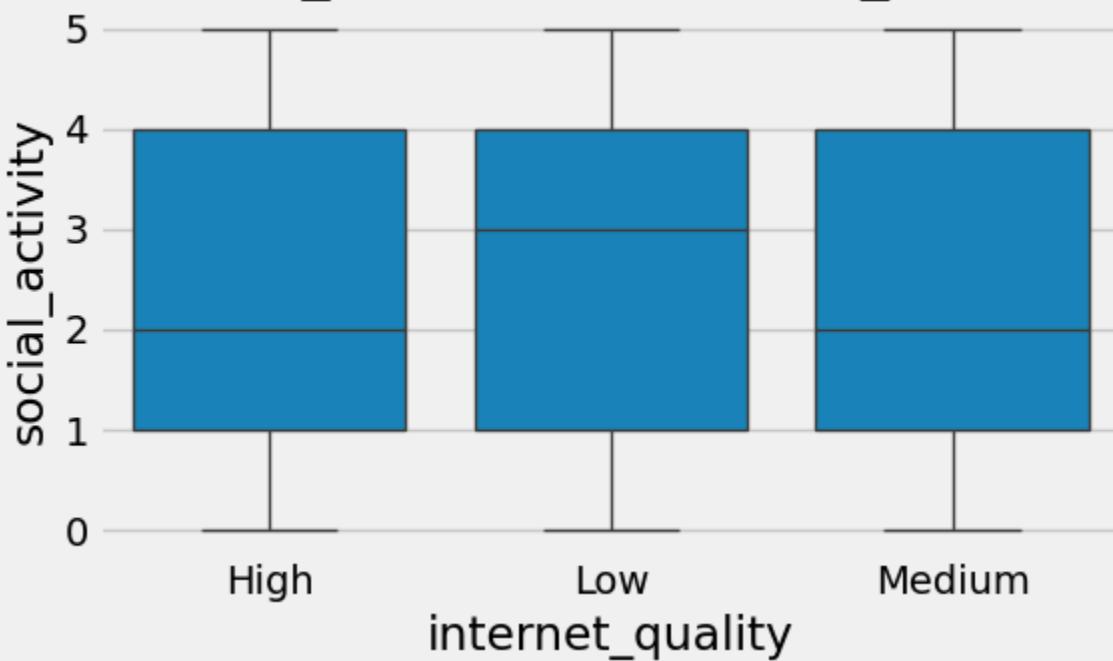




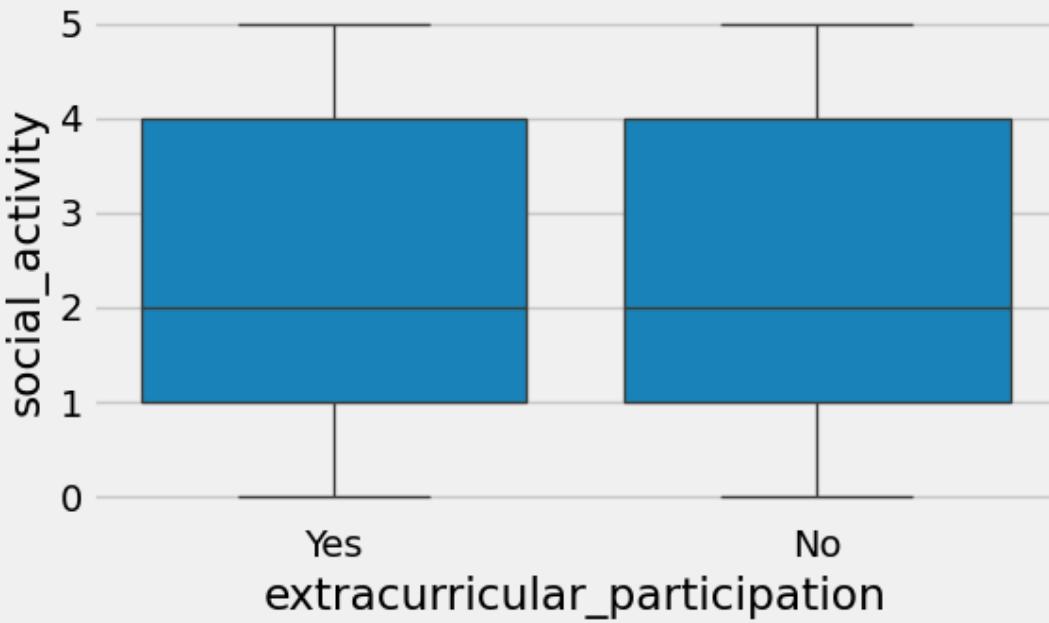
`social_activity` by `parental_education_level`



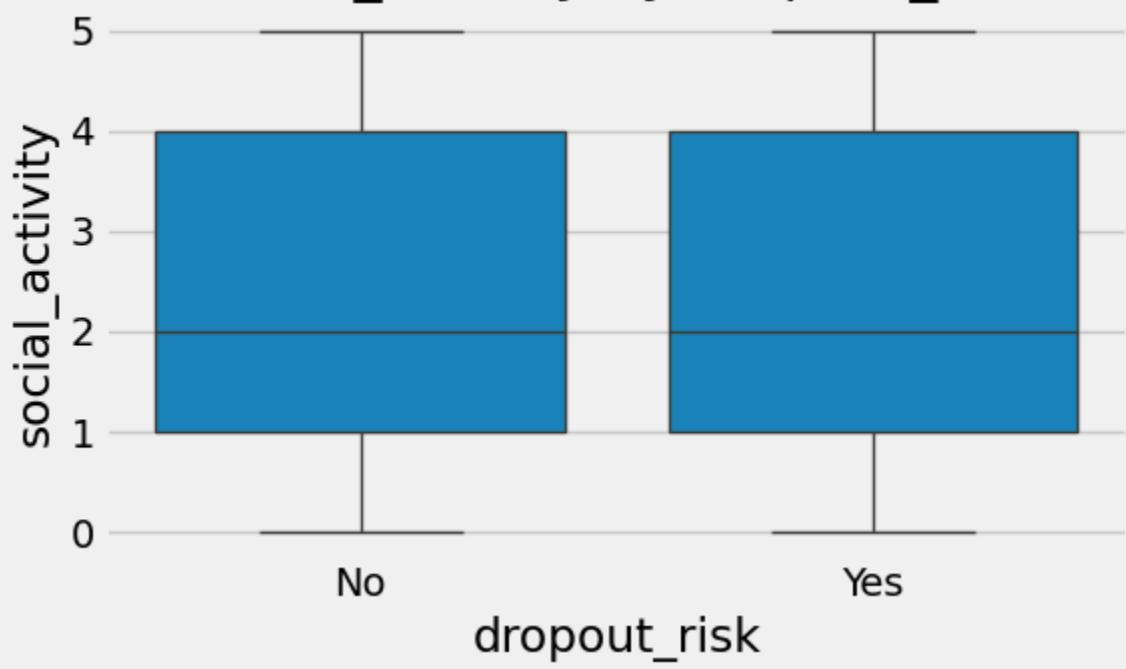
`social_activity` by `internet_quality`

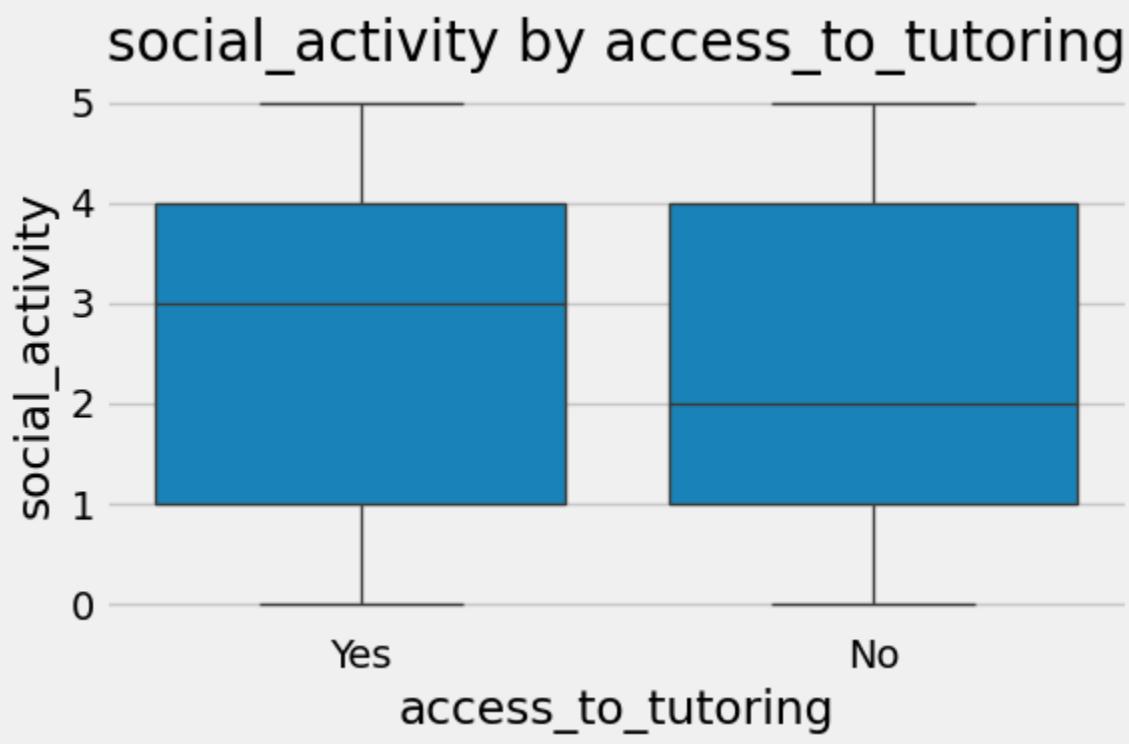
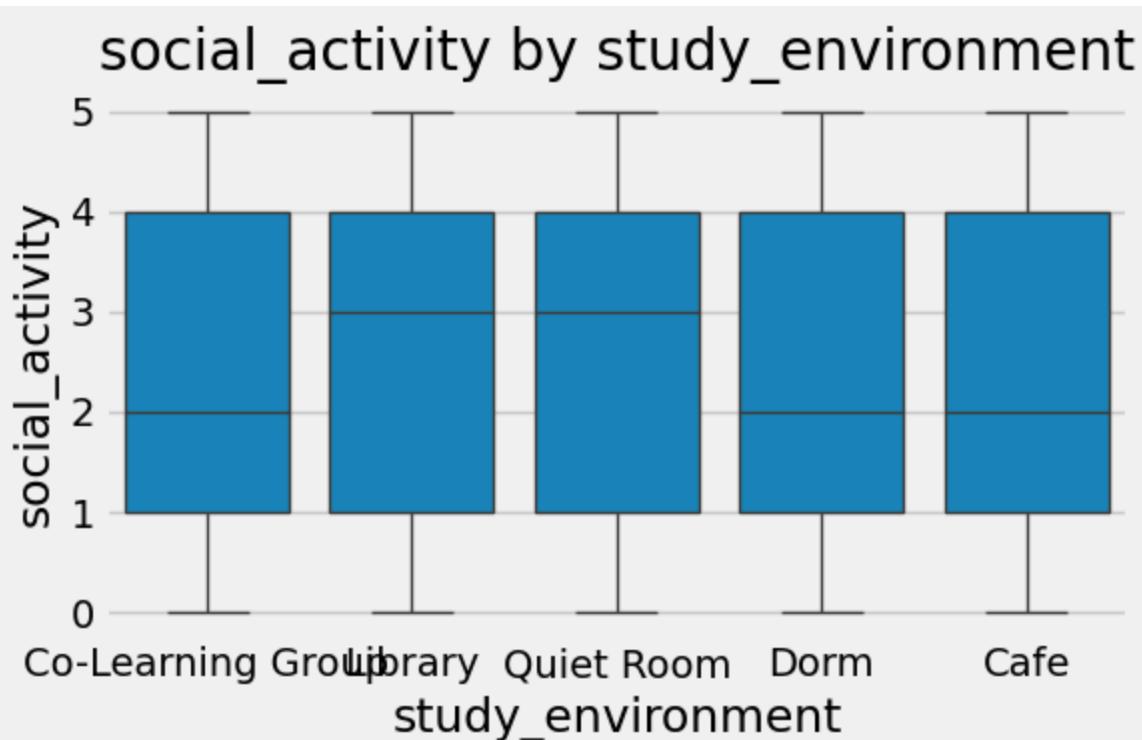


social_activity by extracurricular_participation

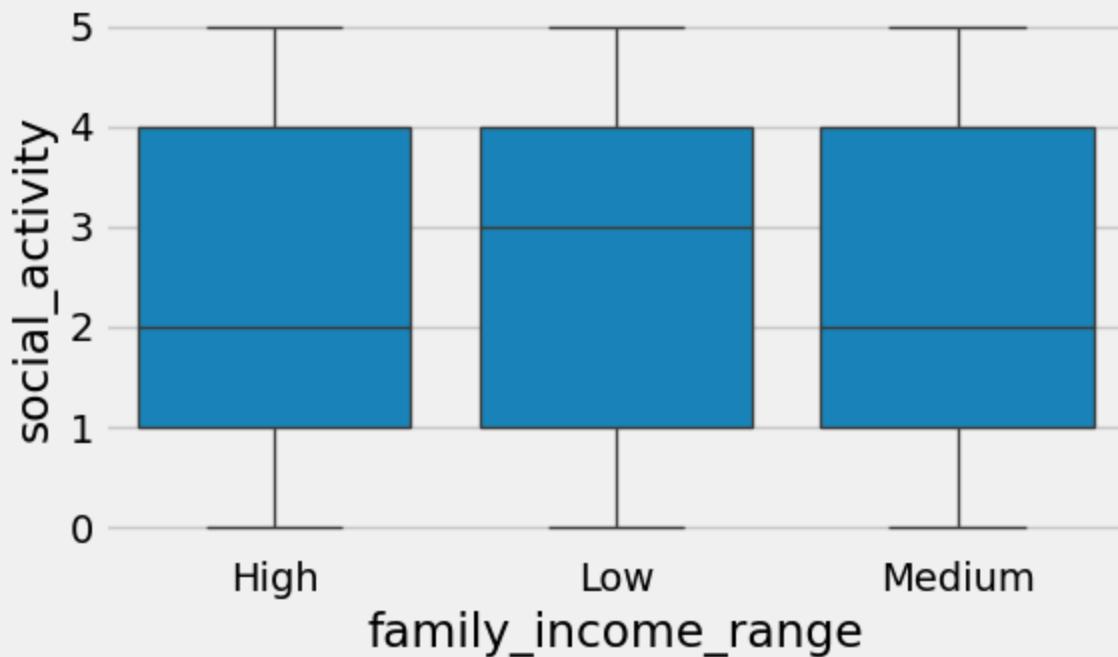


social_activity by dropout_risk

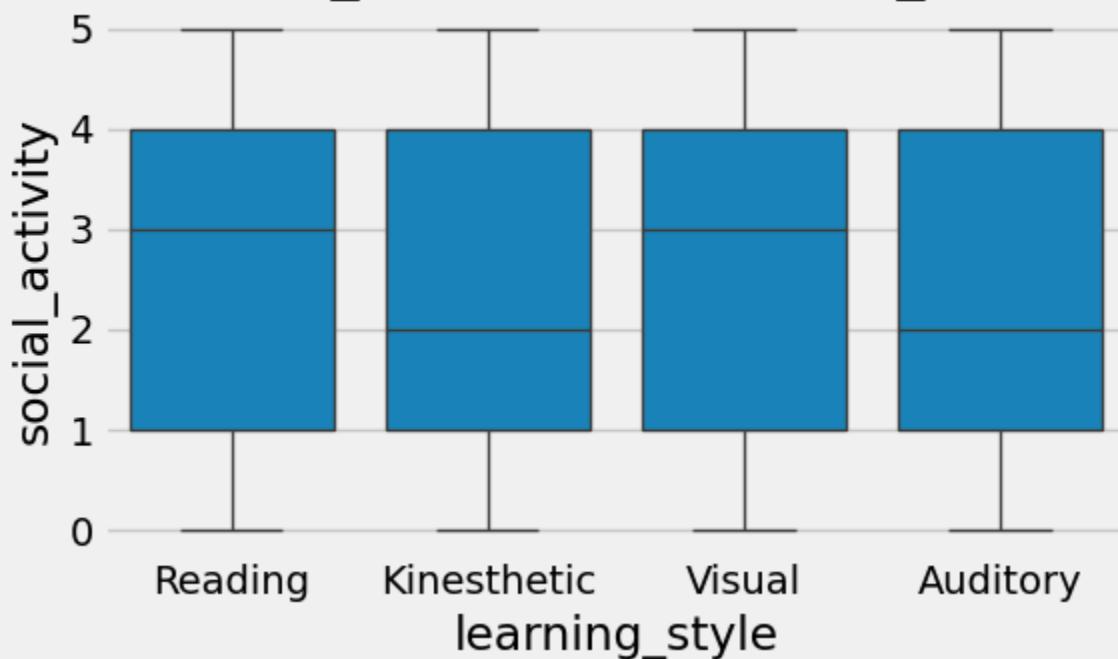




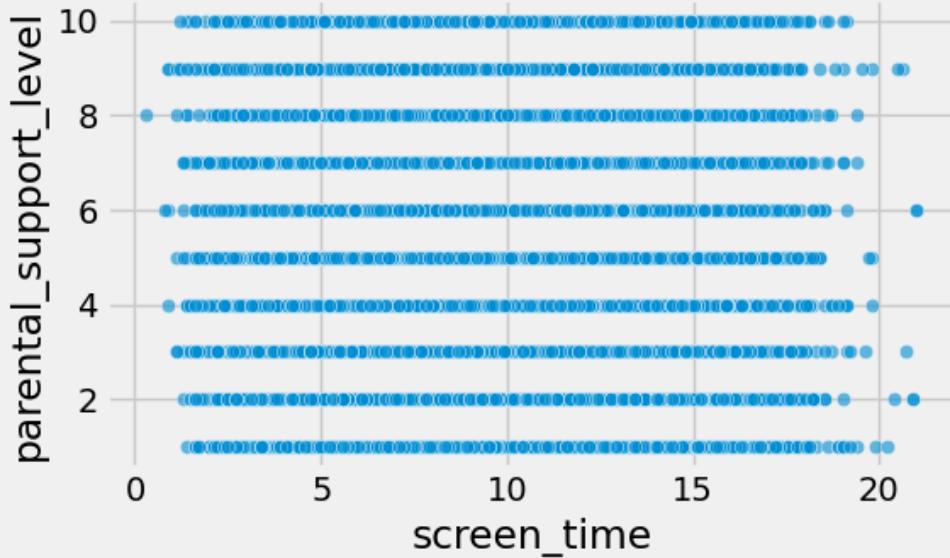
social_activity by family_income_range



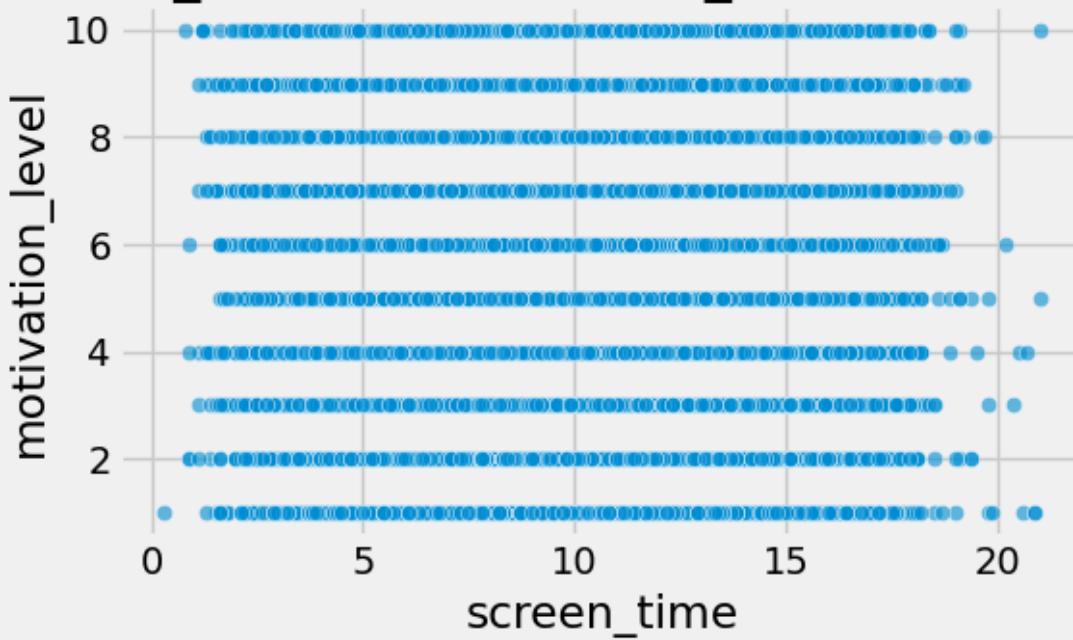
social_activity by learning_style



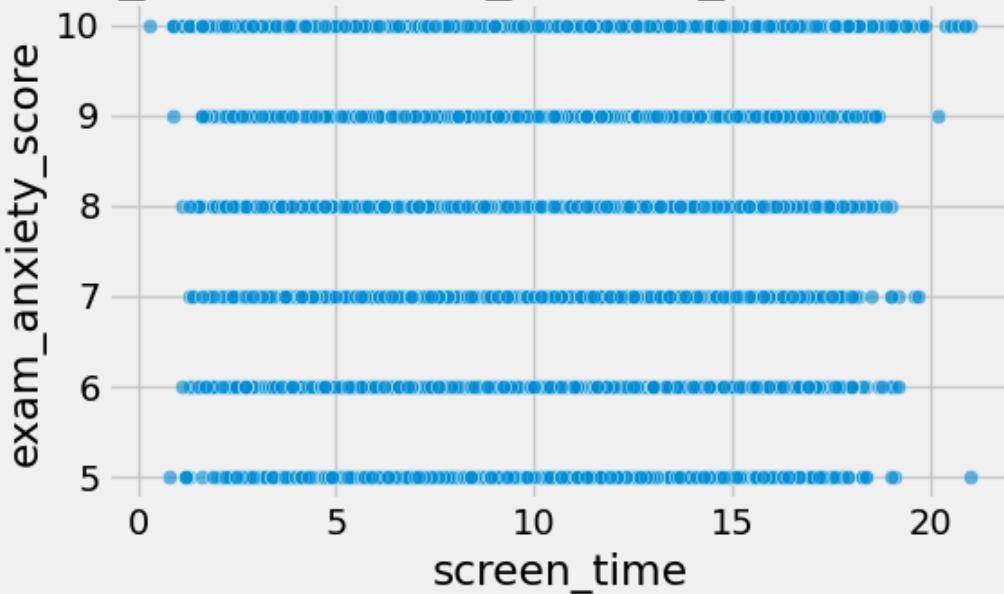
screen_time vs parental_support_level (corr=-0.00)



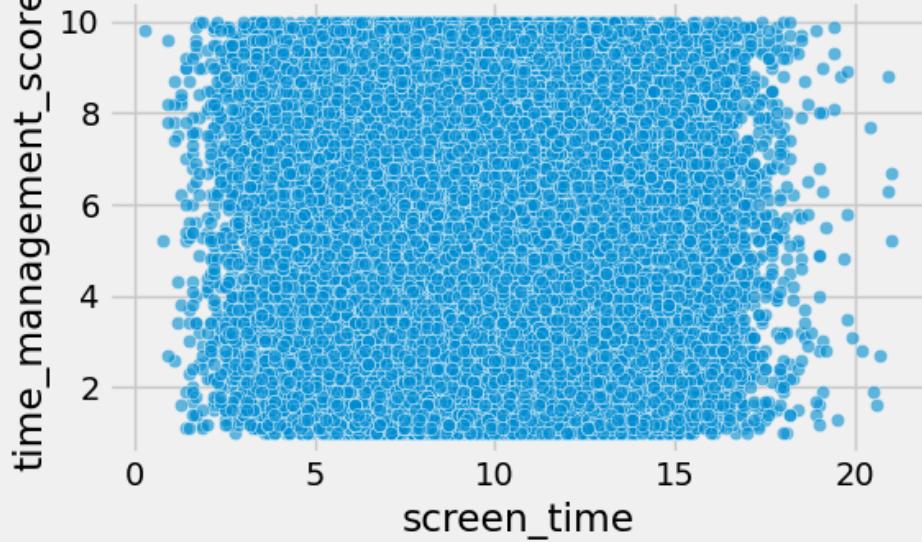
screen_time vs motivation_level (corr=-0.00)

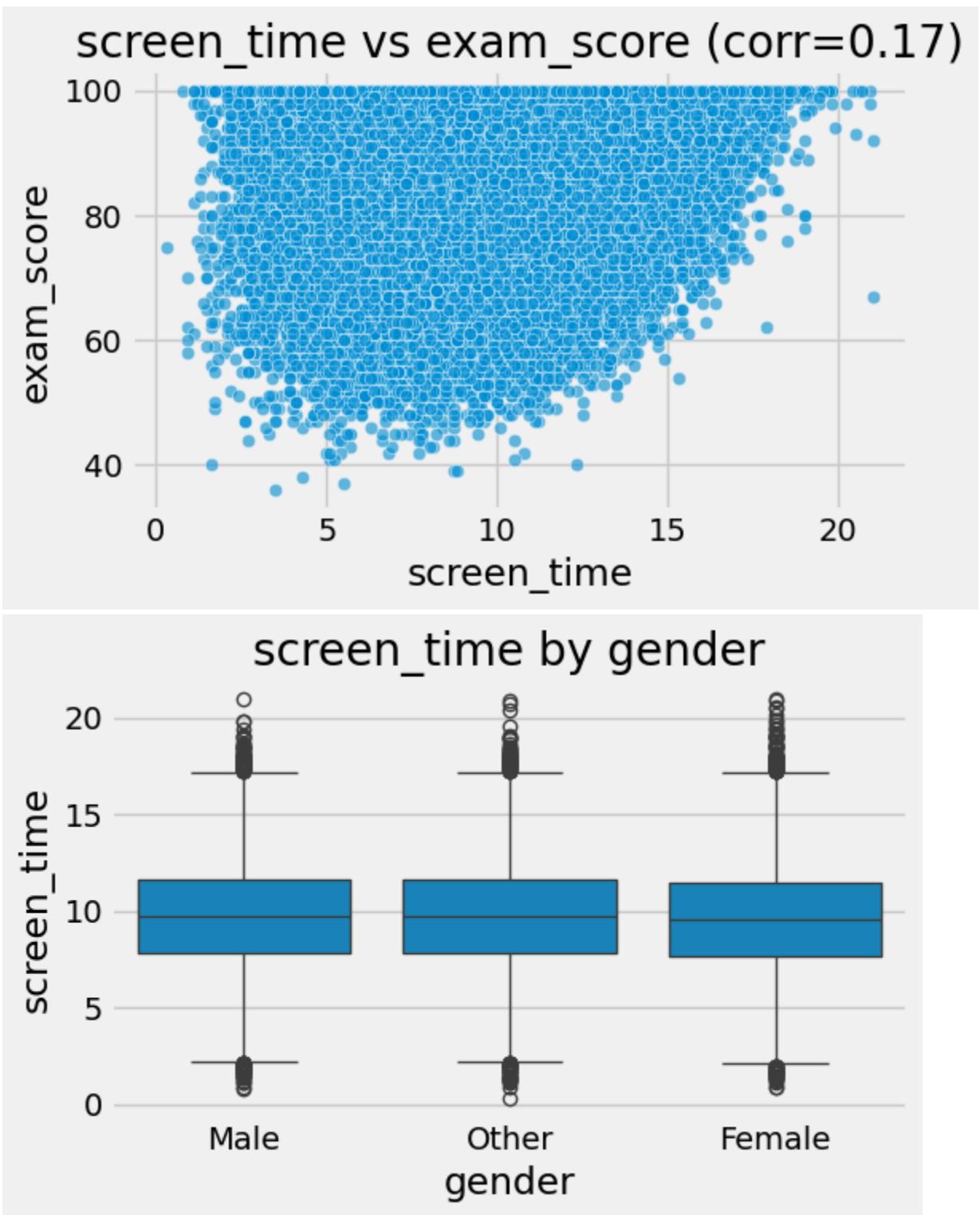


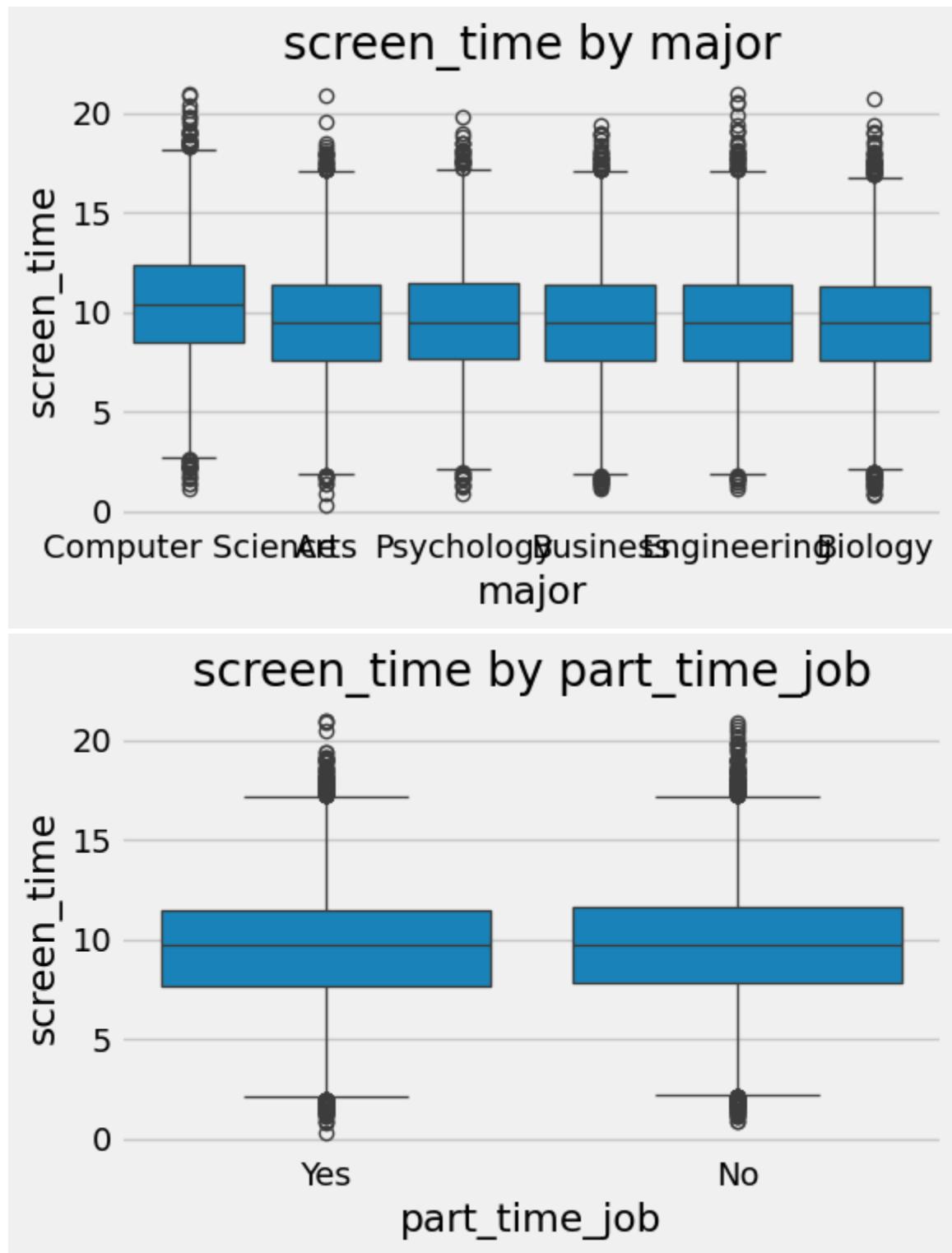
screen_time vs exam_anxiety_score (corr=0.00)

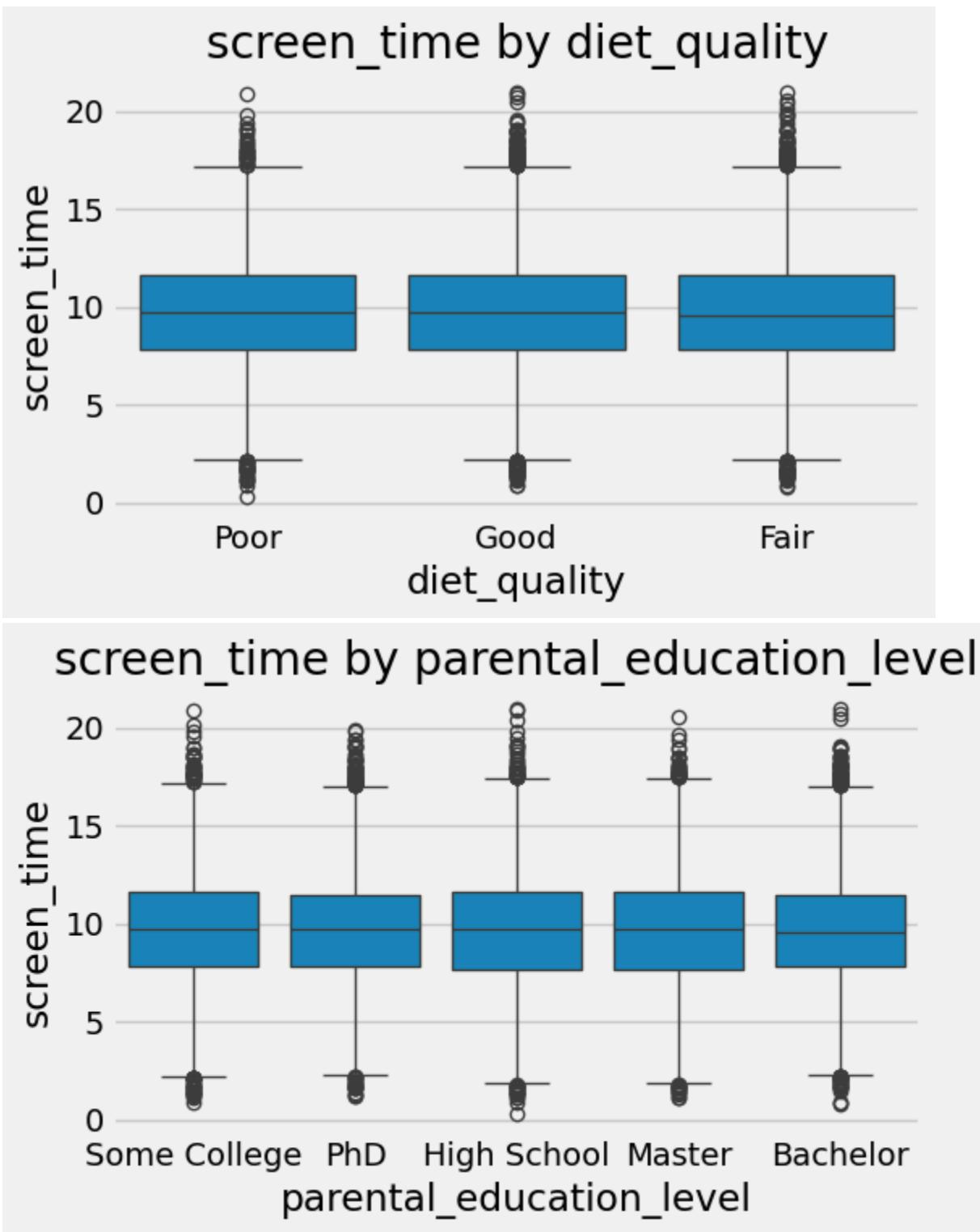


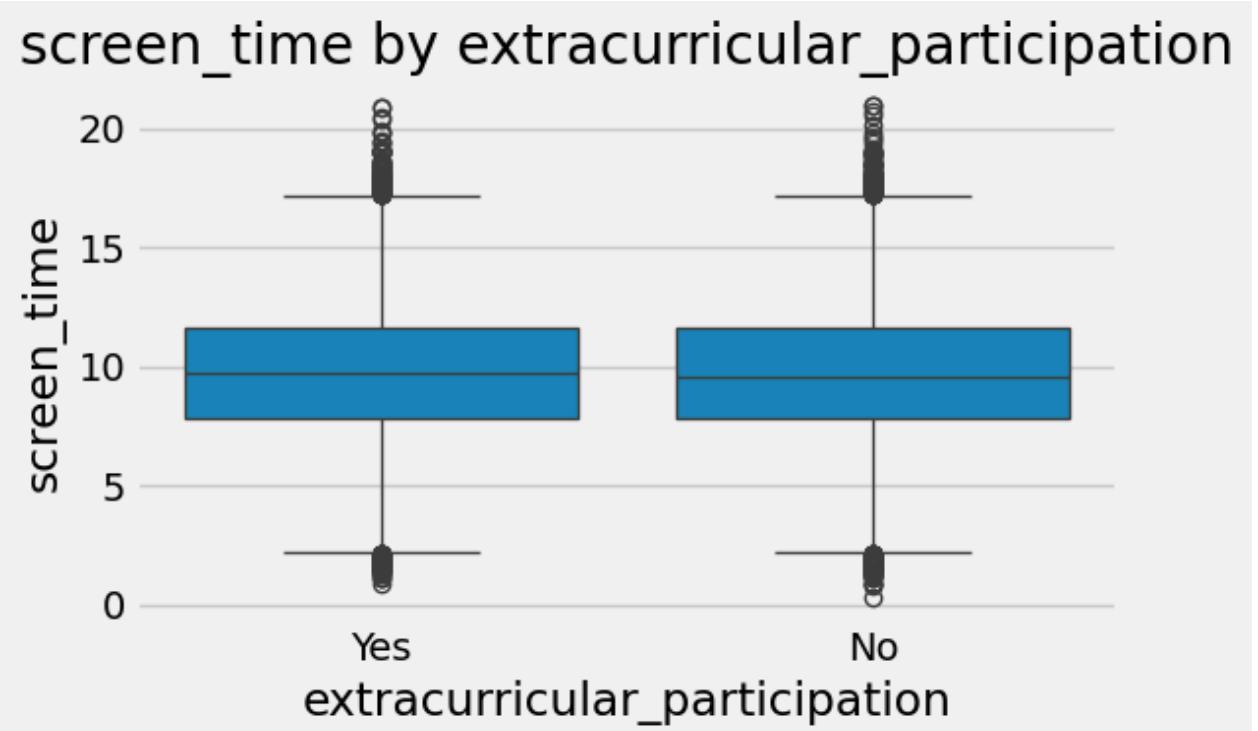
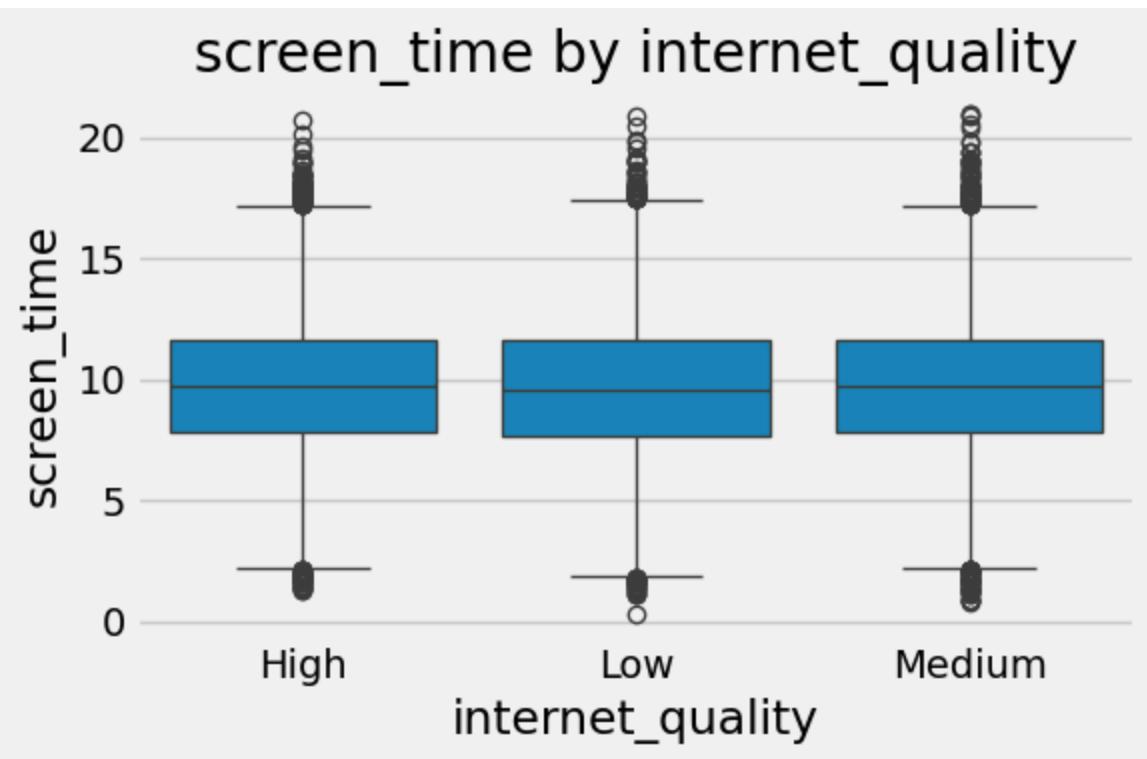
screen_time vs time_management_score (corr=0.01)

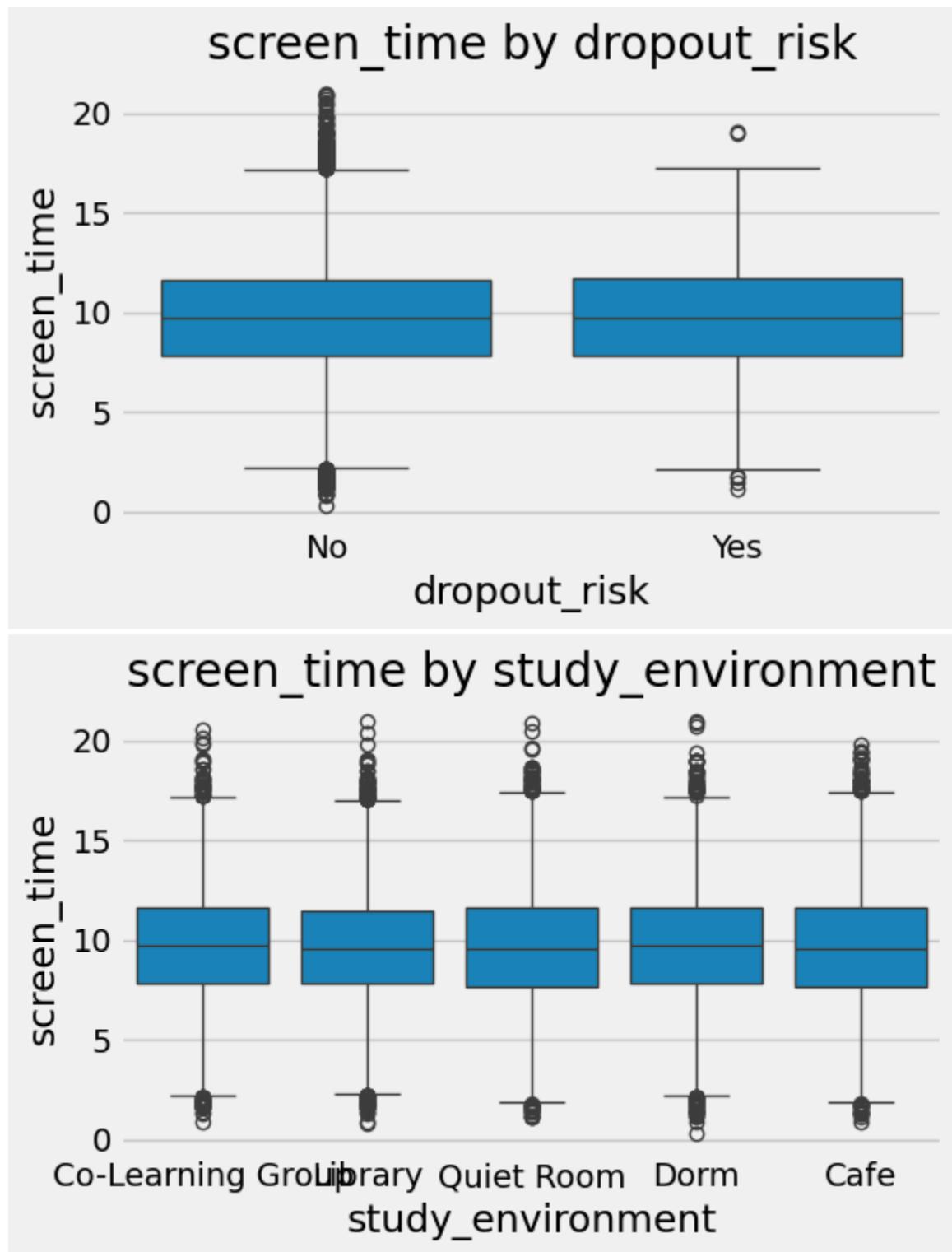


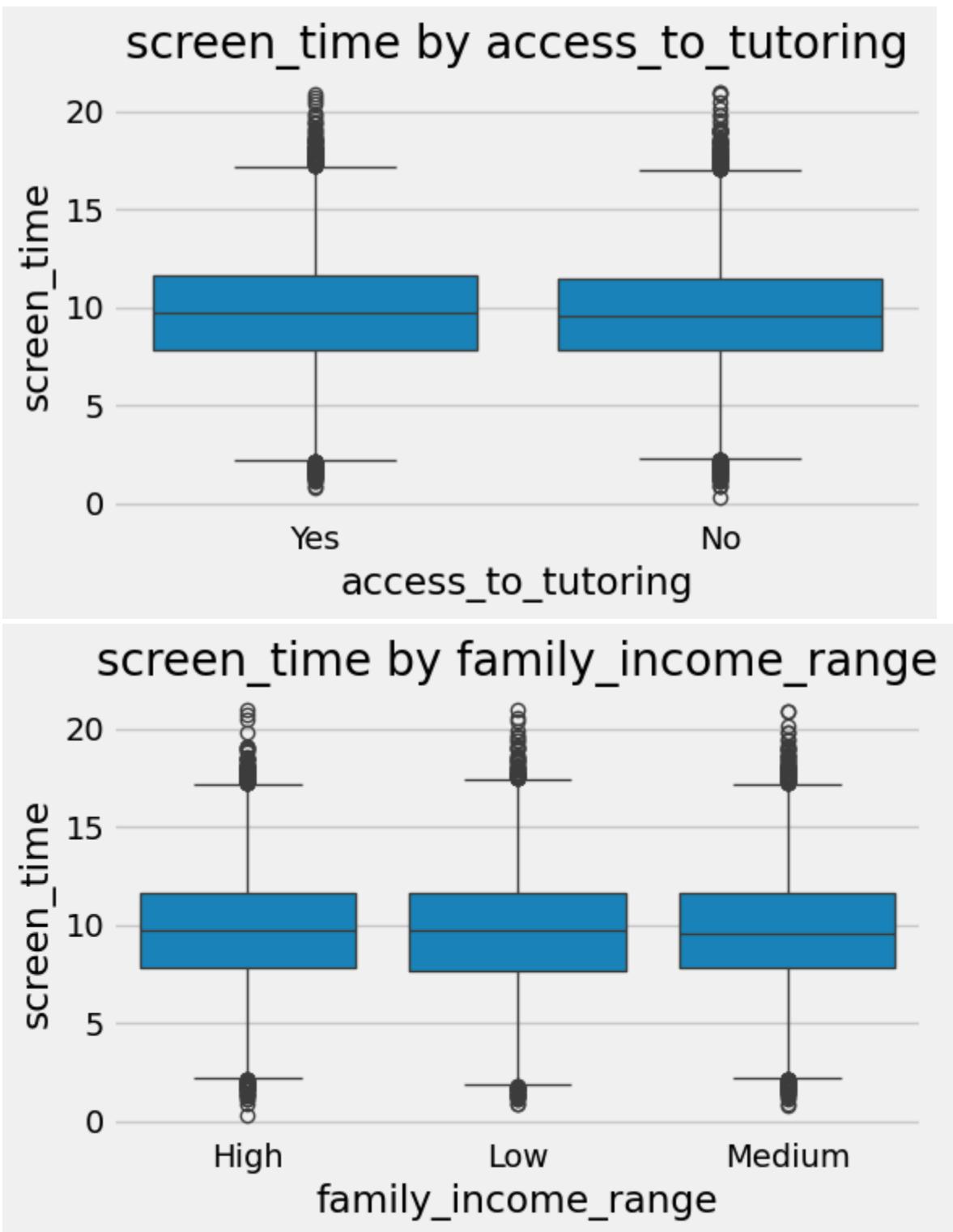


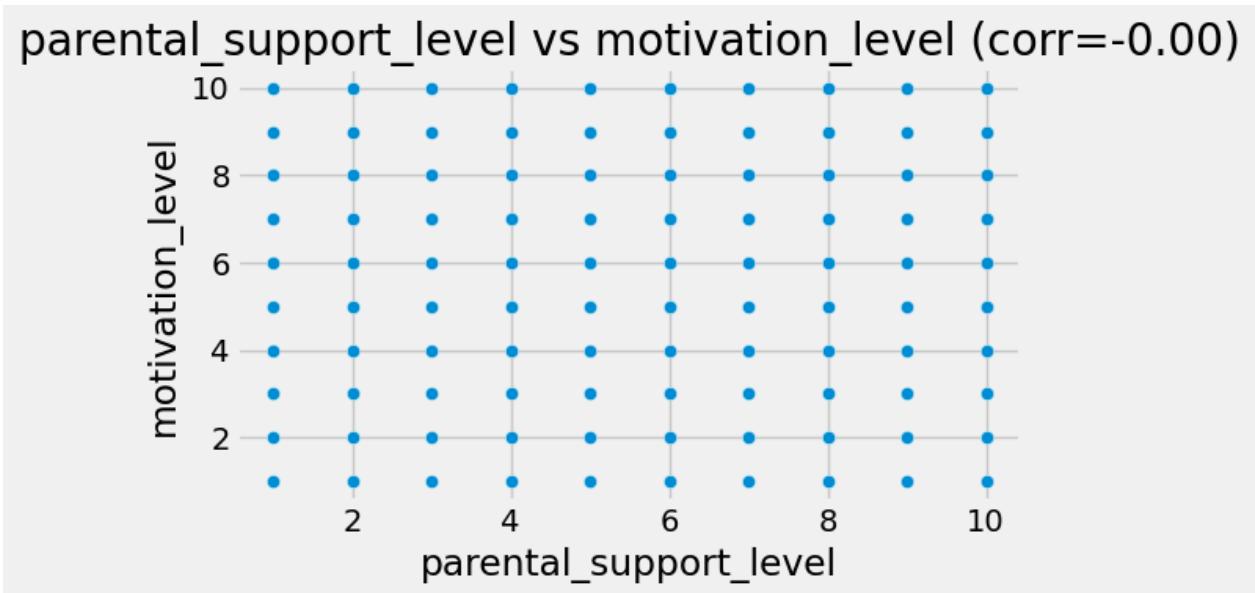
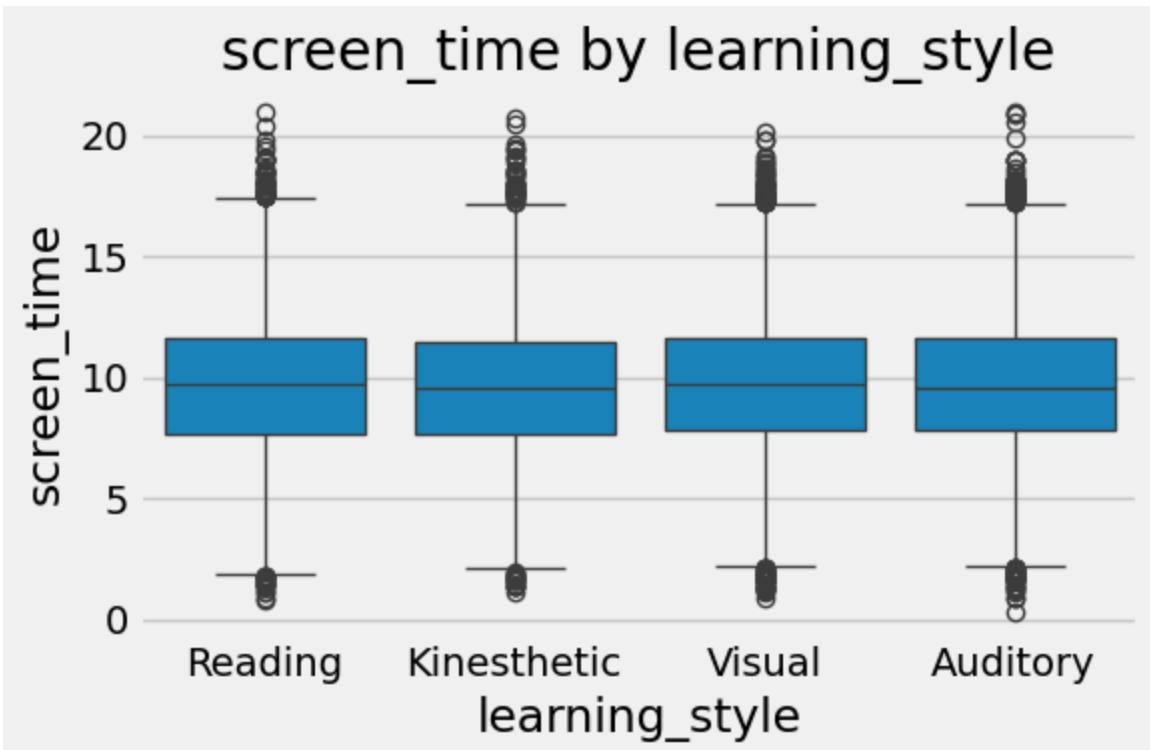




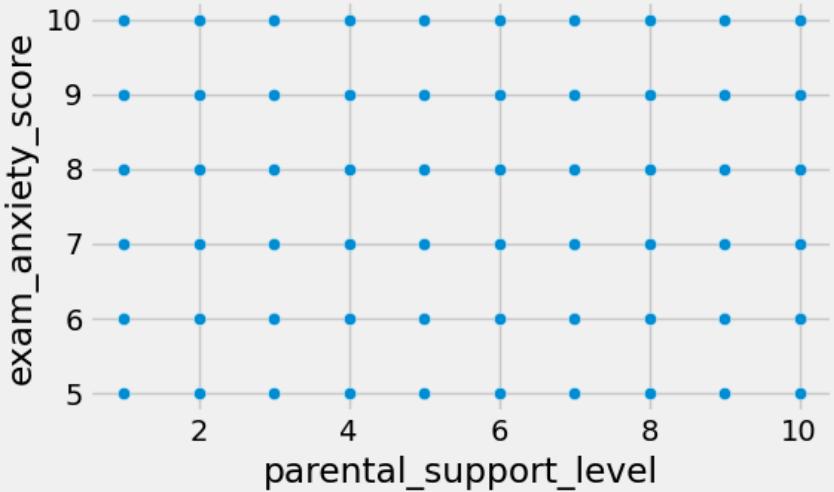




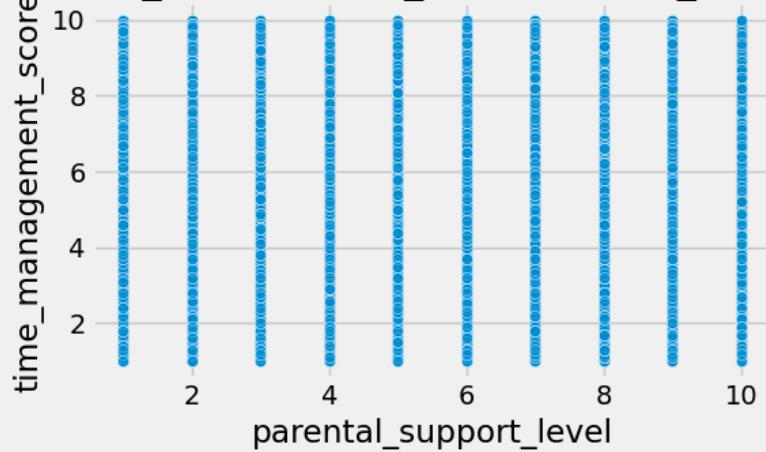




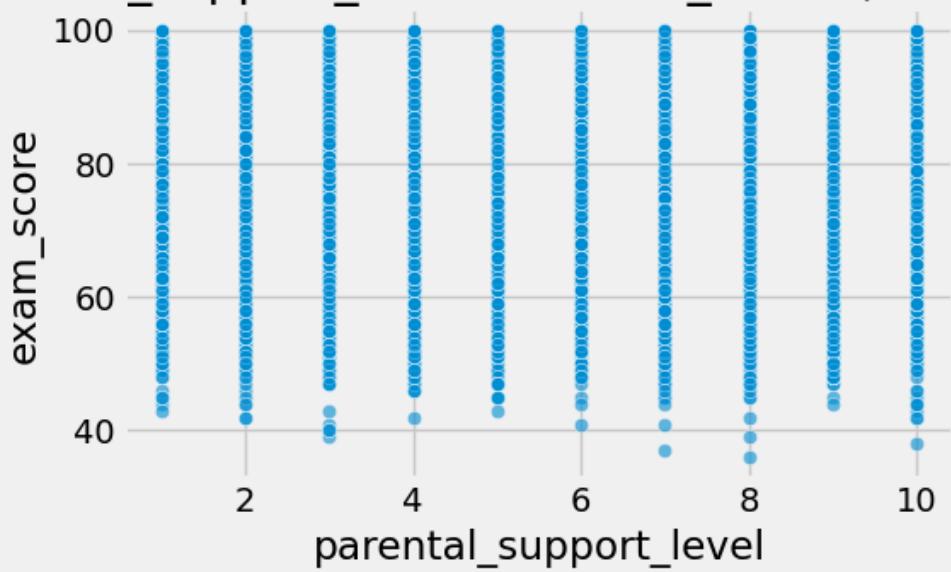
parental_support_level vs exam_anxiety_score (corr=0.00)

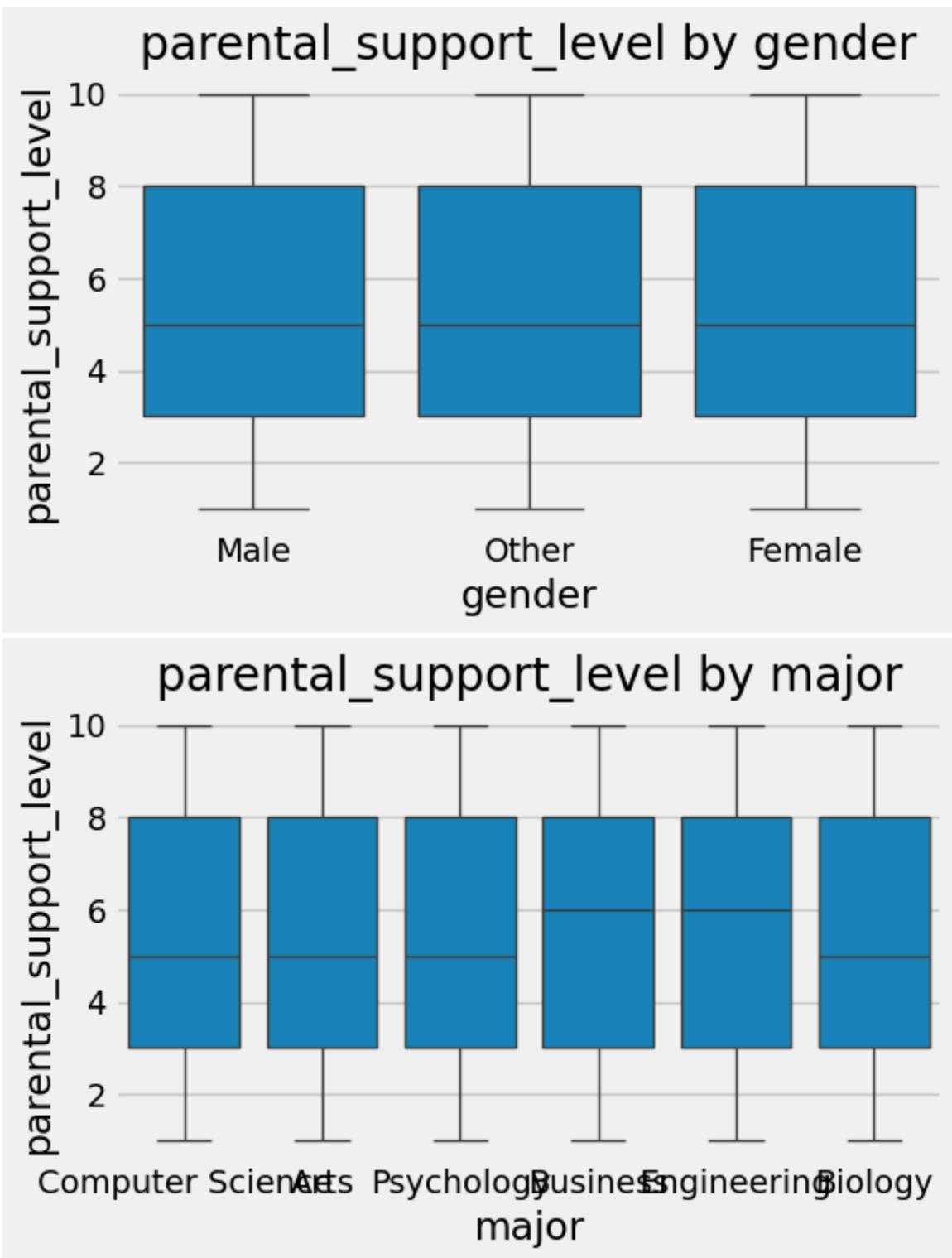


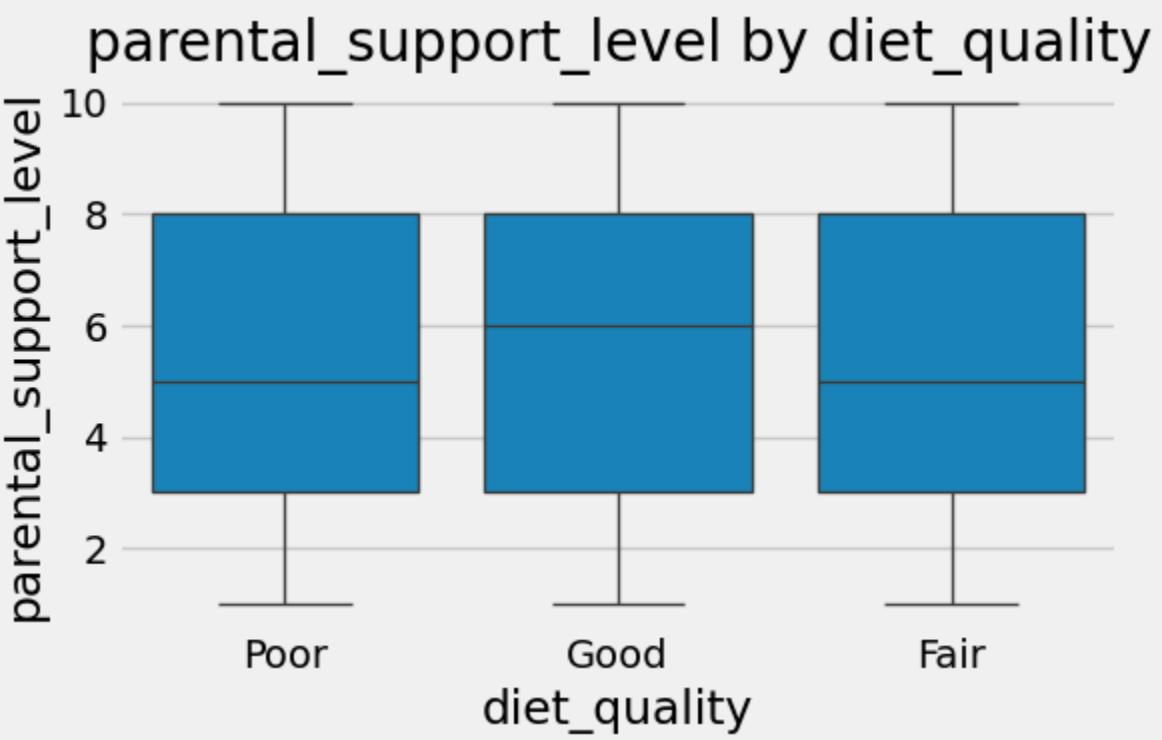
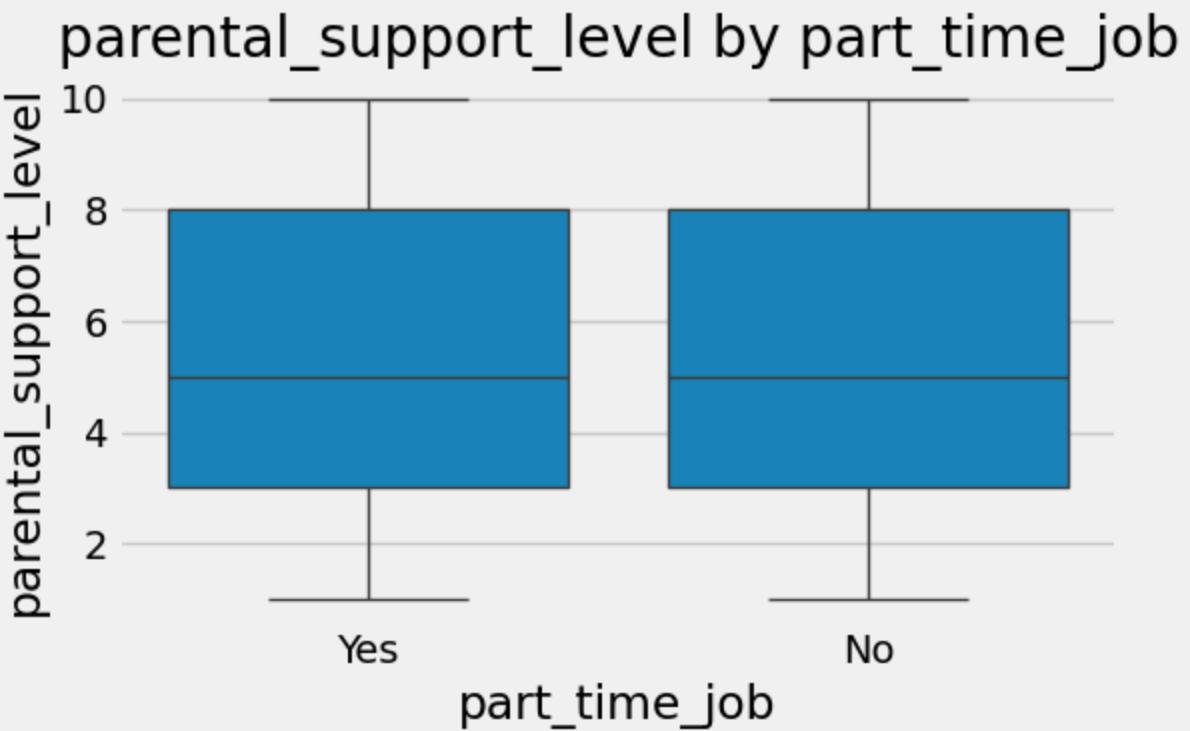
parental_support_level vs time_management_score (corr=-0.00)

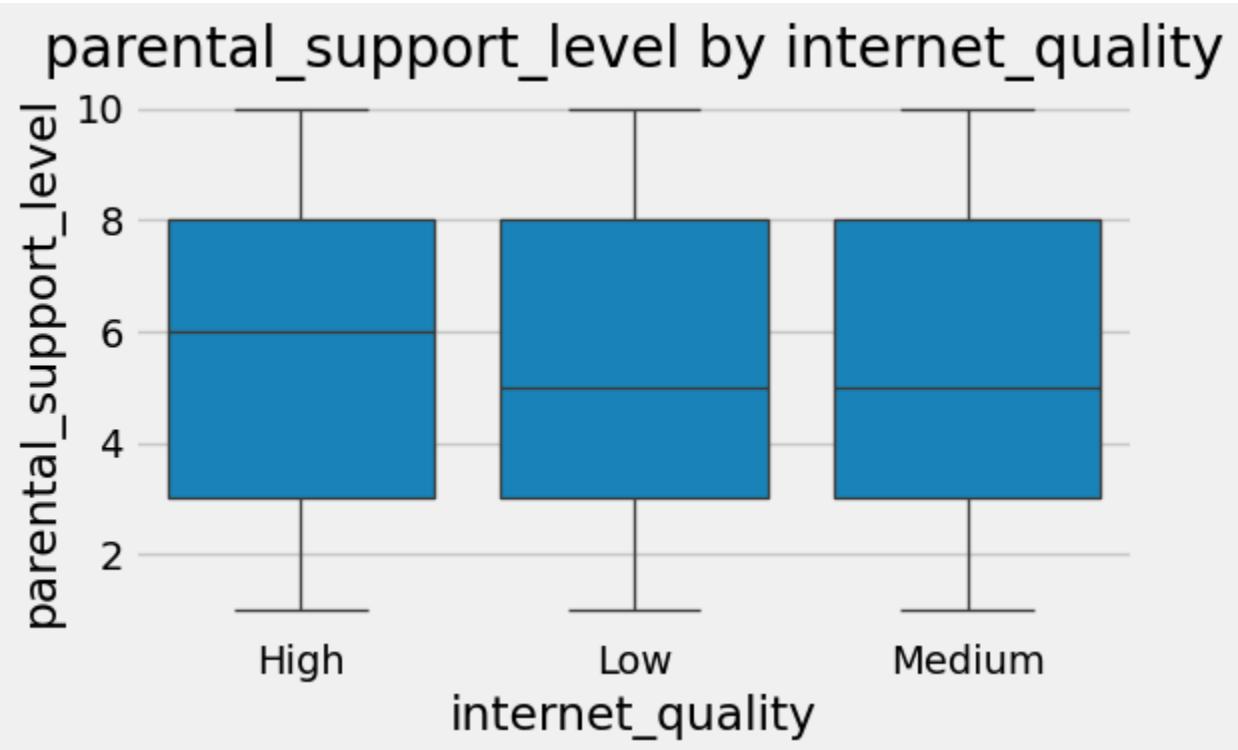
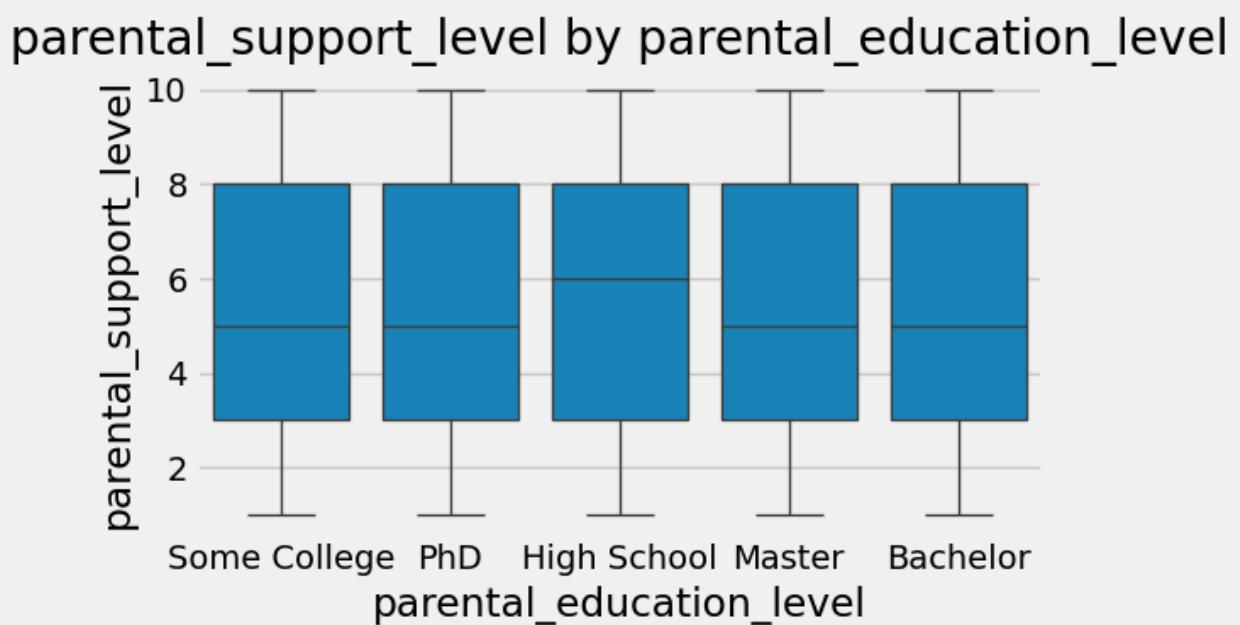


parental_support_level vs exam_score (corr=-0.01)

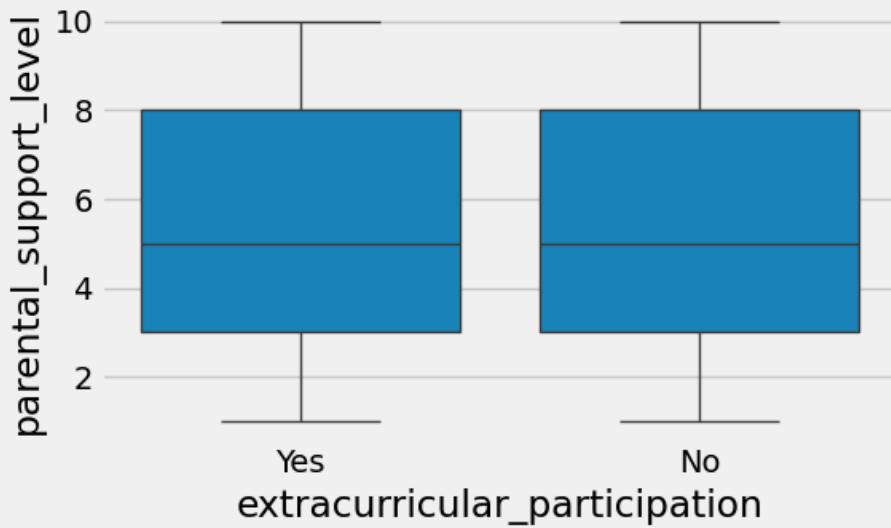




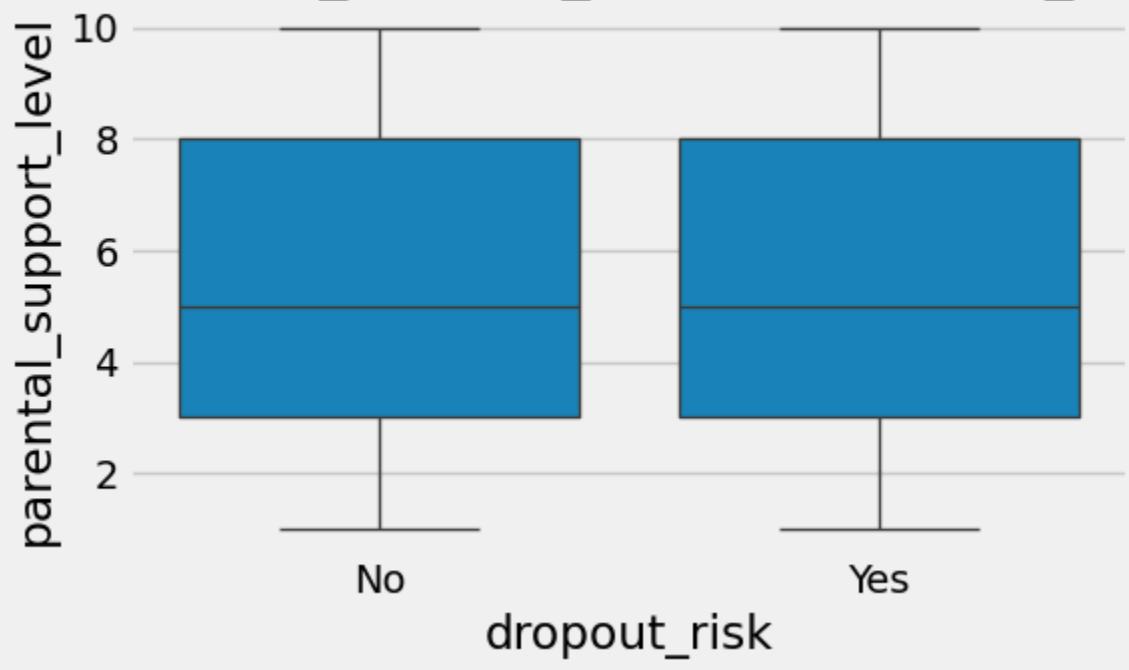




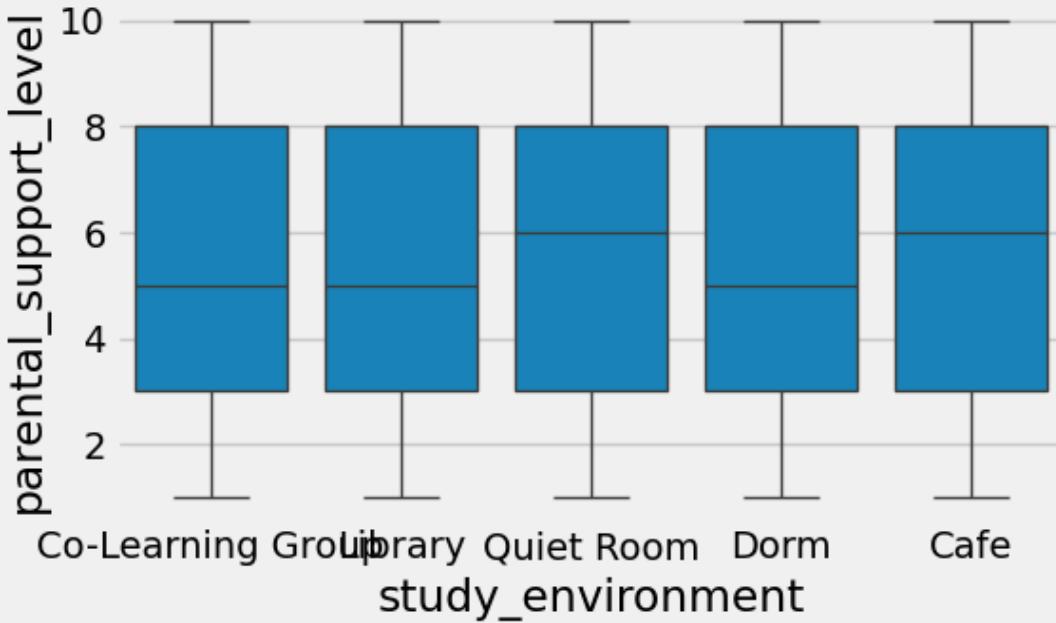
parental_support_level by extracurricular_participation



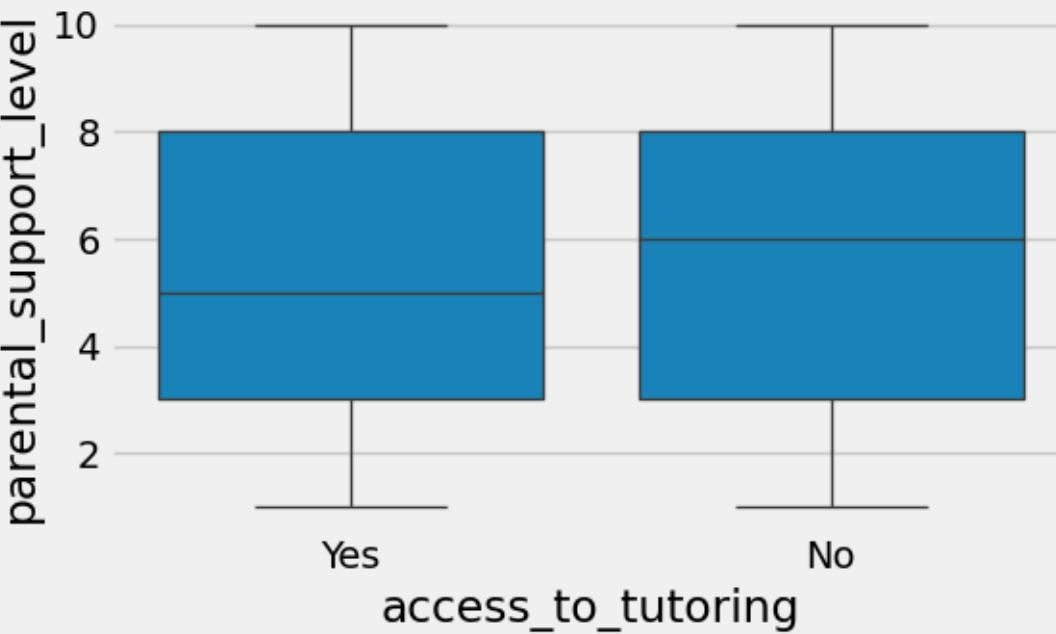
parental_support_level by dropout_risk

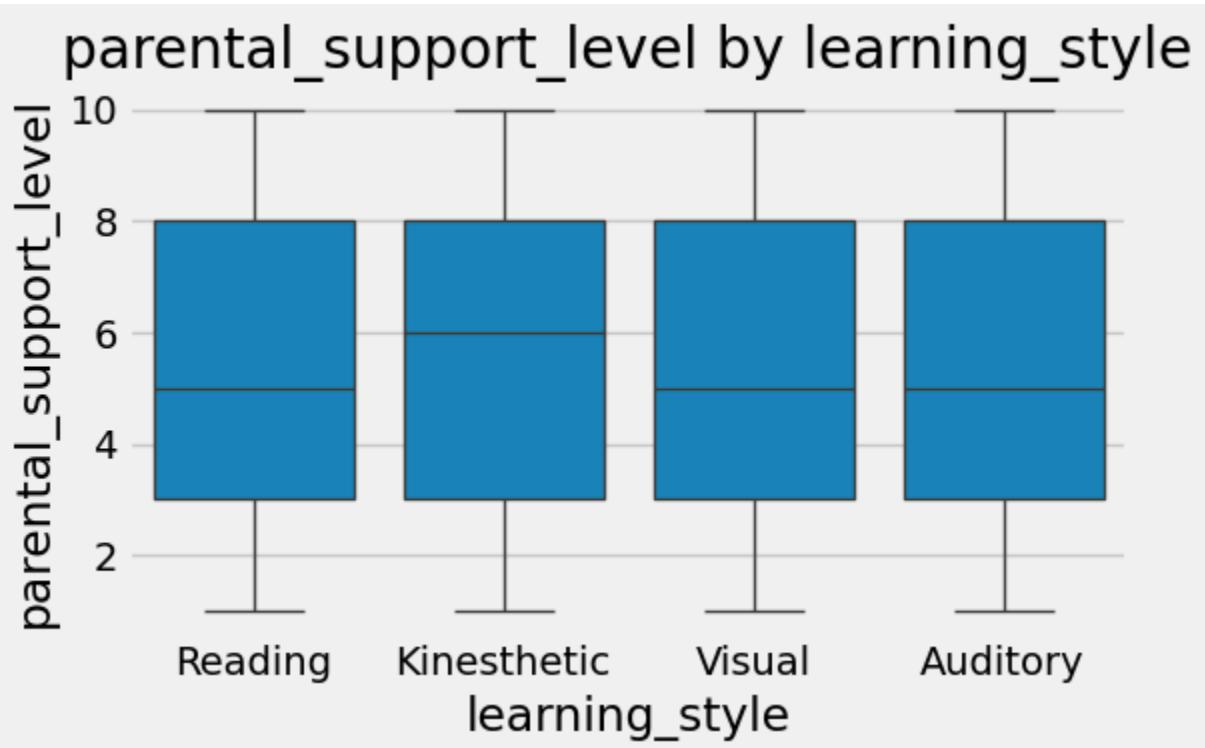
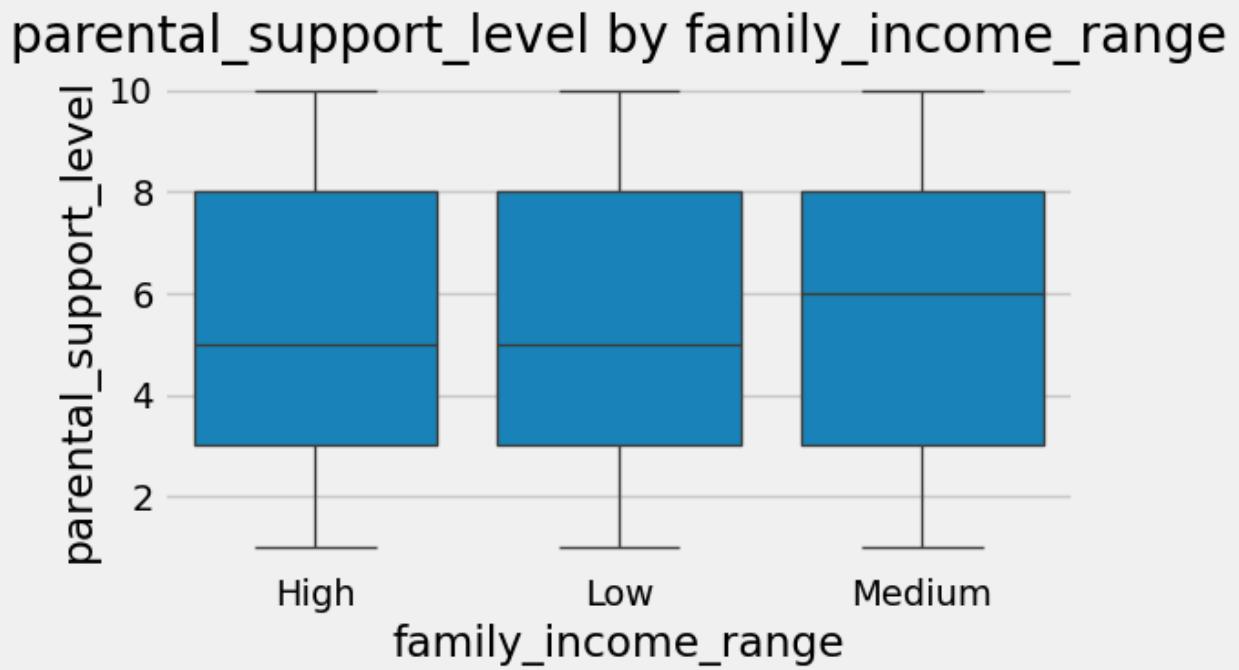


parental_support_level by study_environment

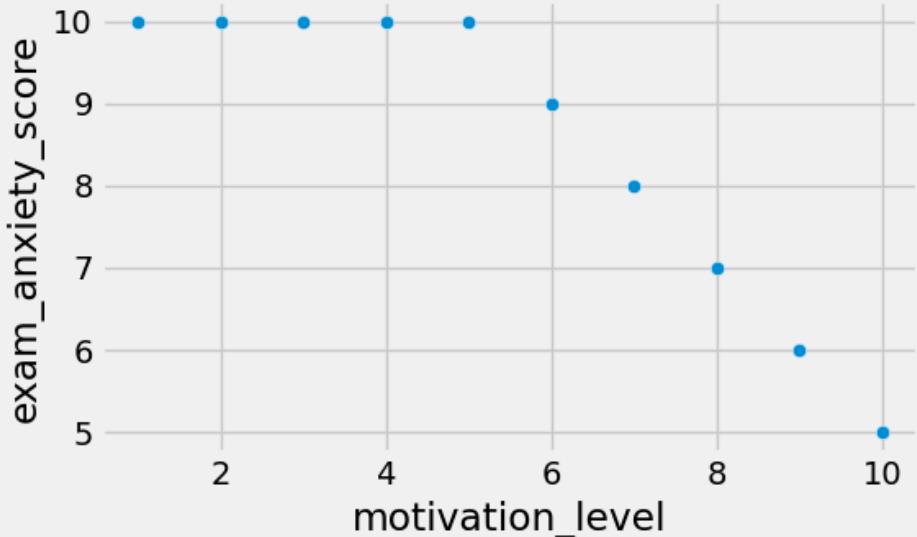


parental_support_level by access_to_tutoring

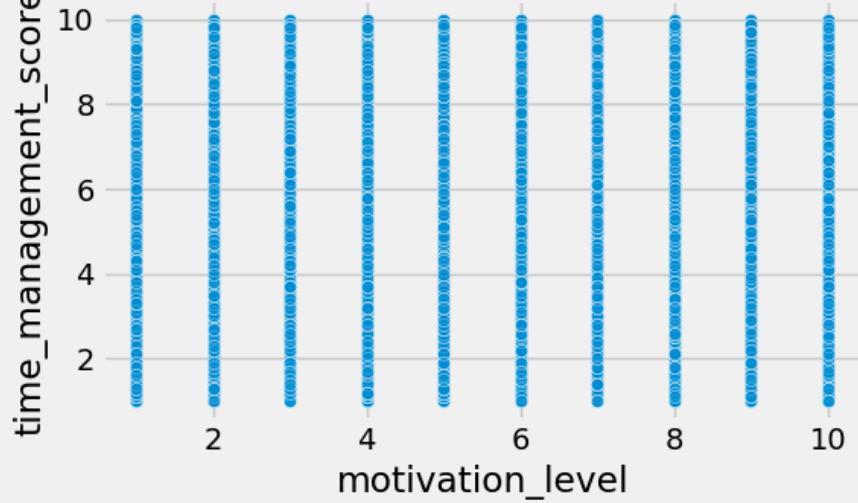


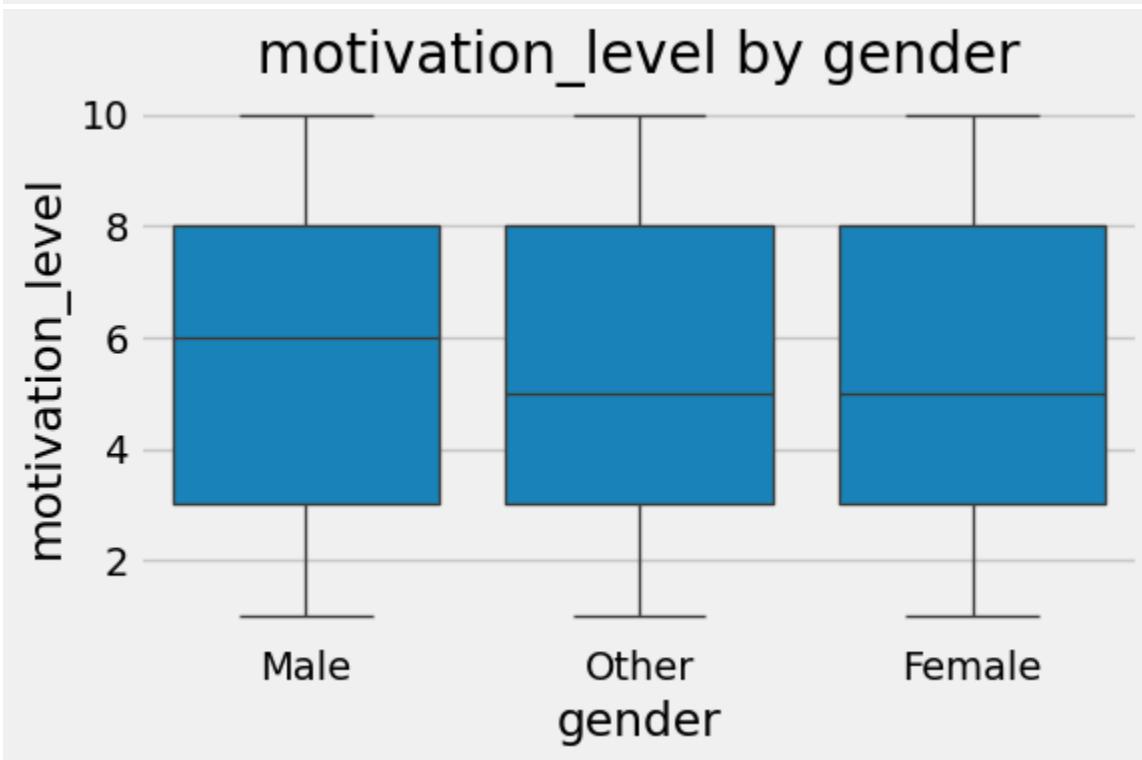
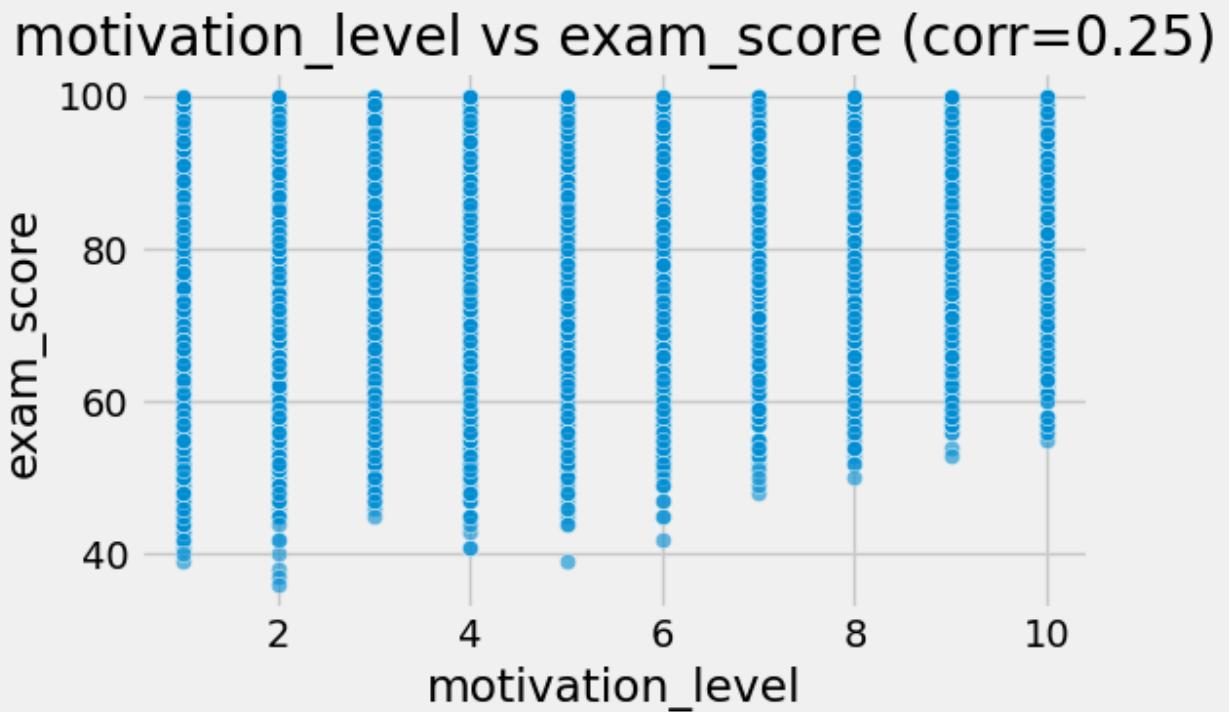


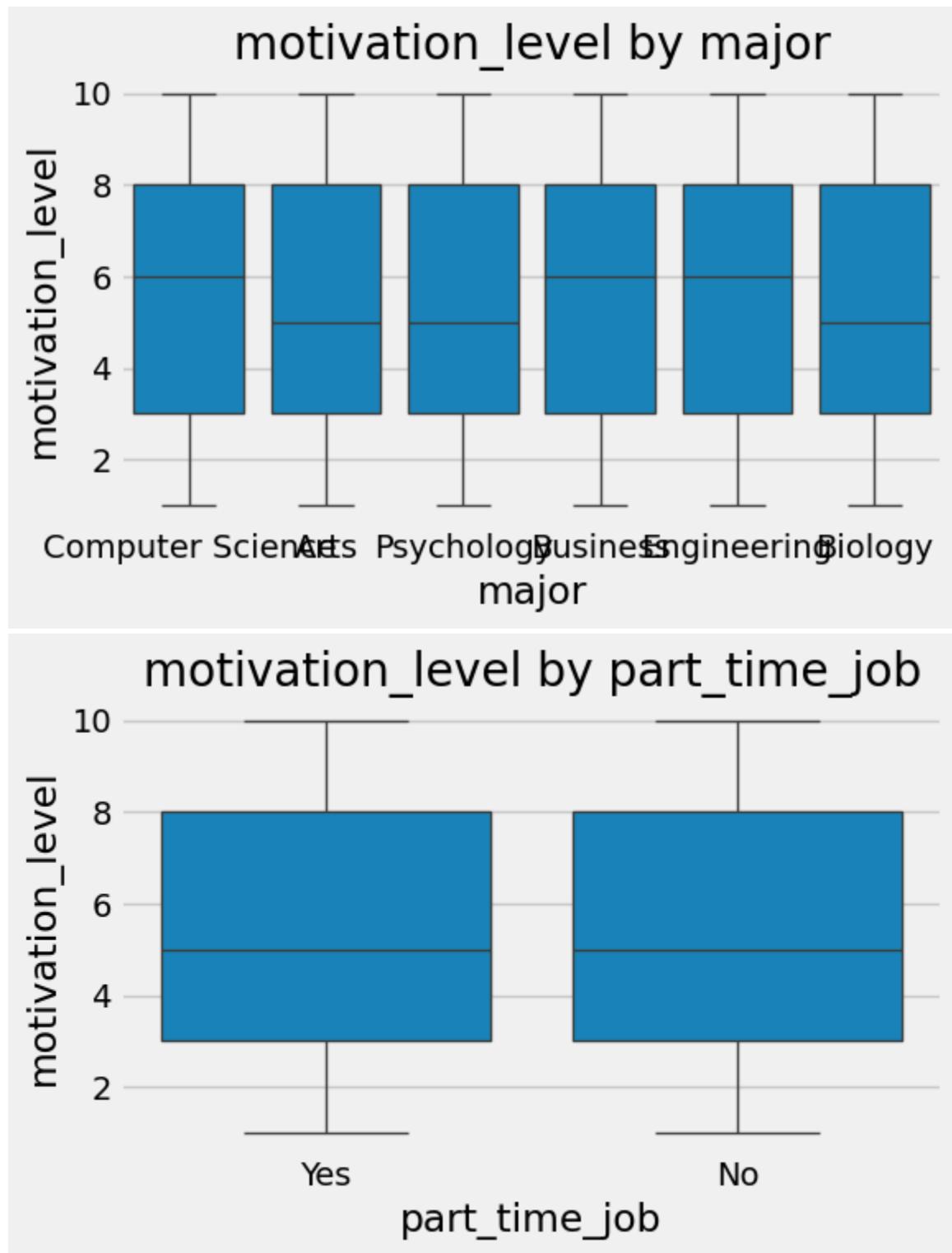
motivation_level vs exam_anxiety_score (corr=-0.92)

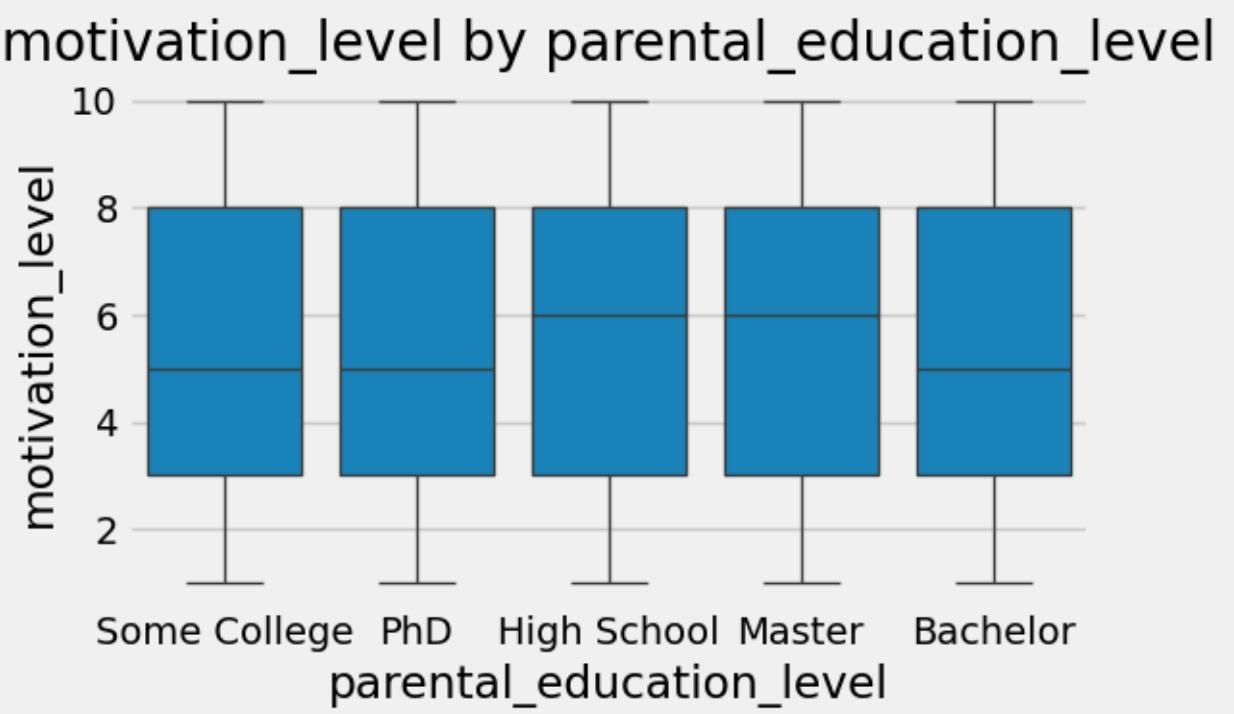
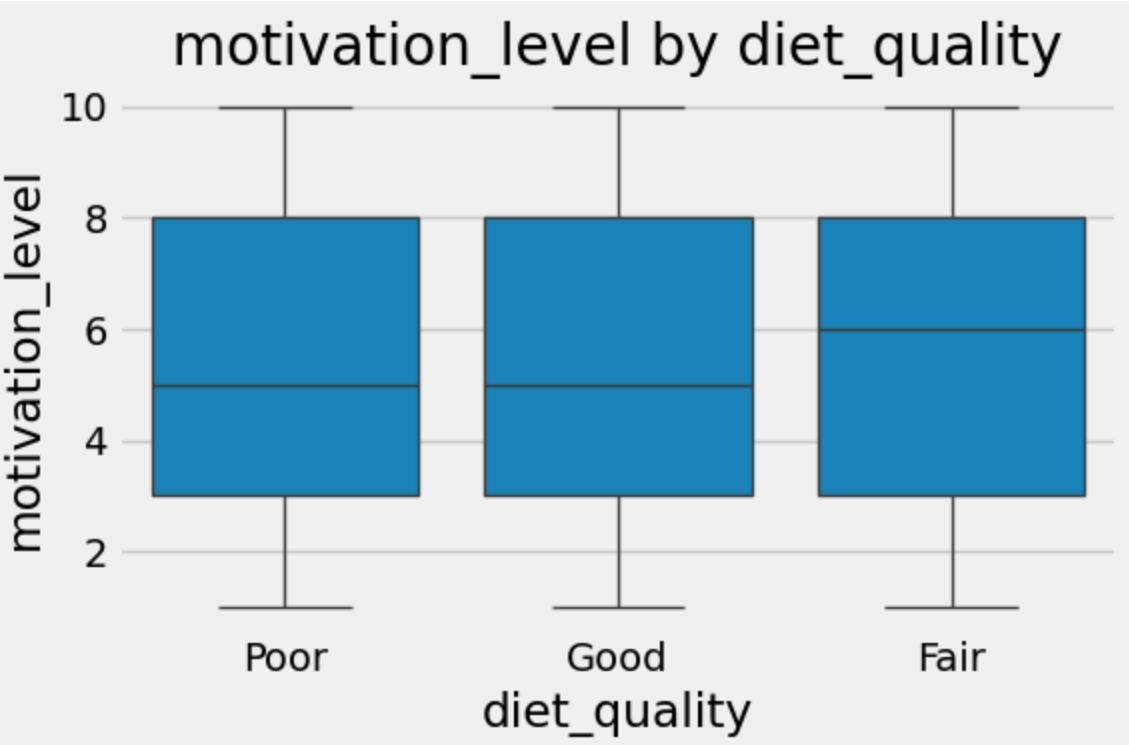


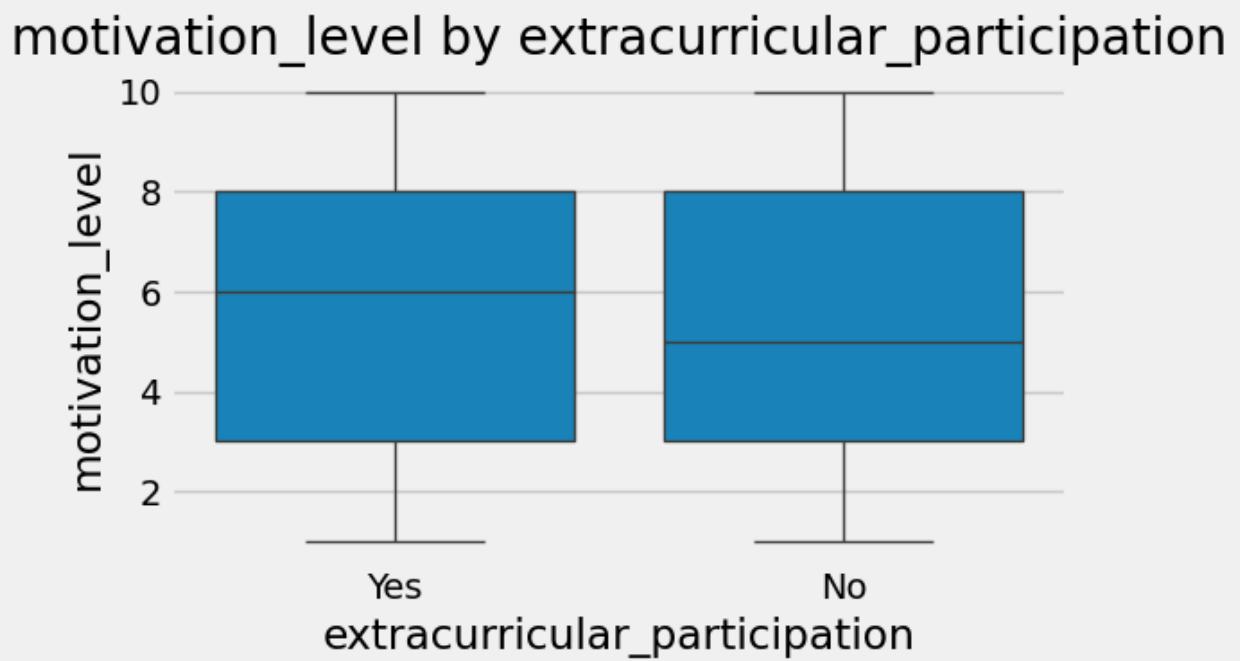
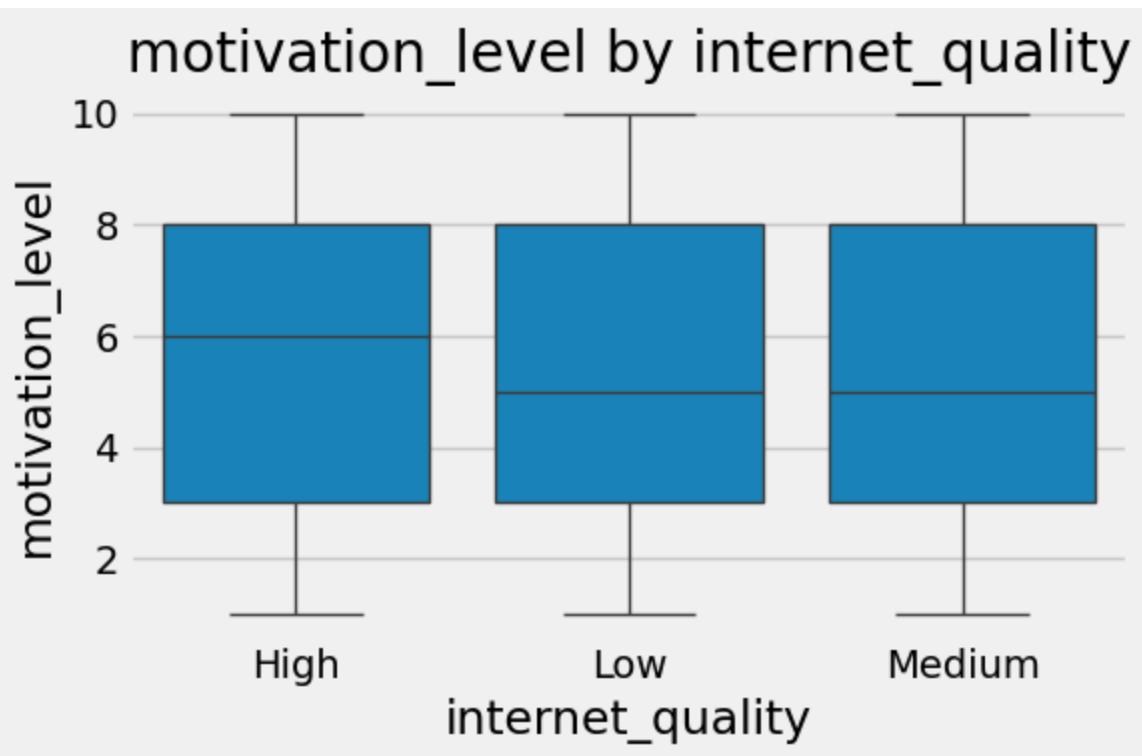
motivation_level vs time_management_score (corr=0.00)

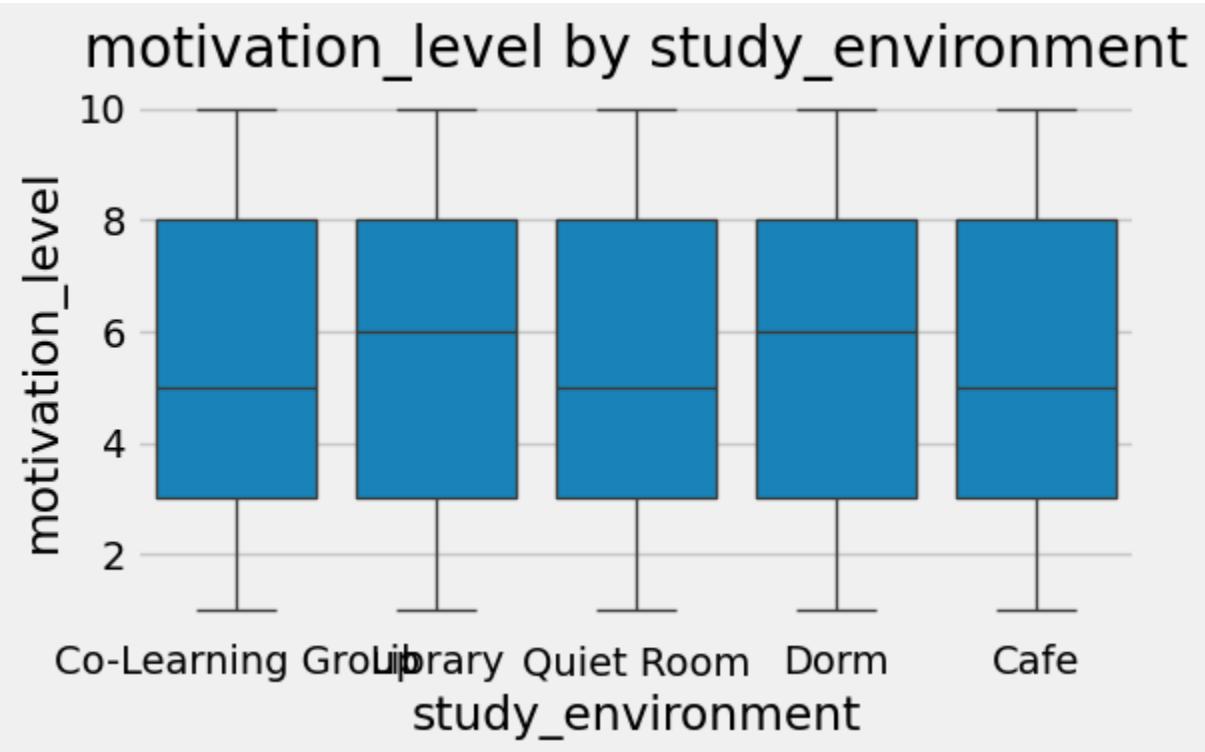
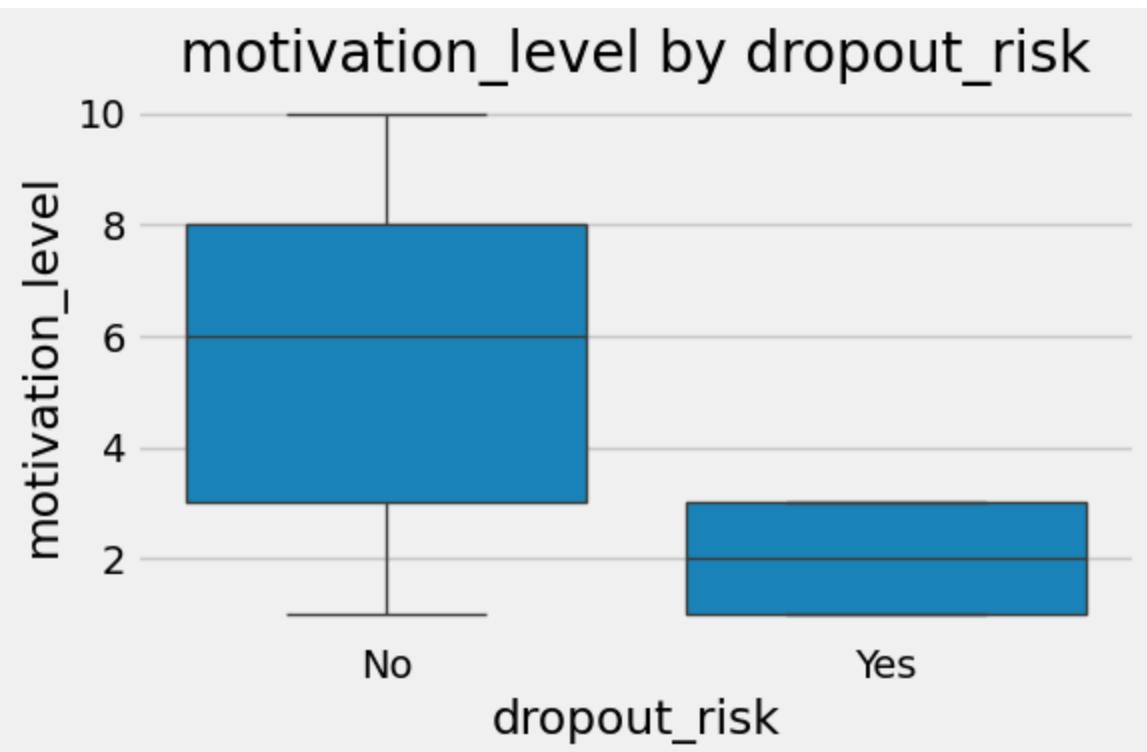


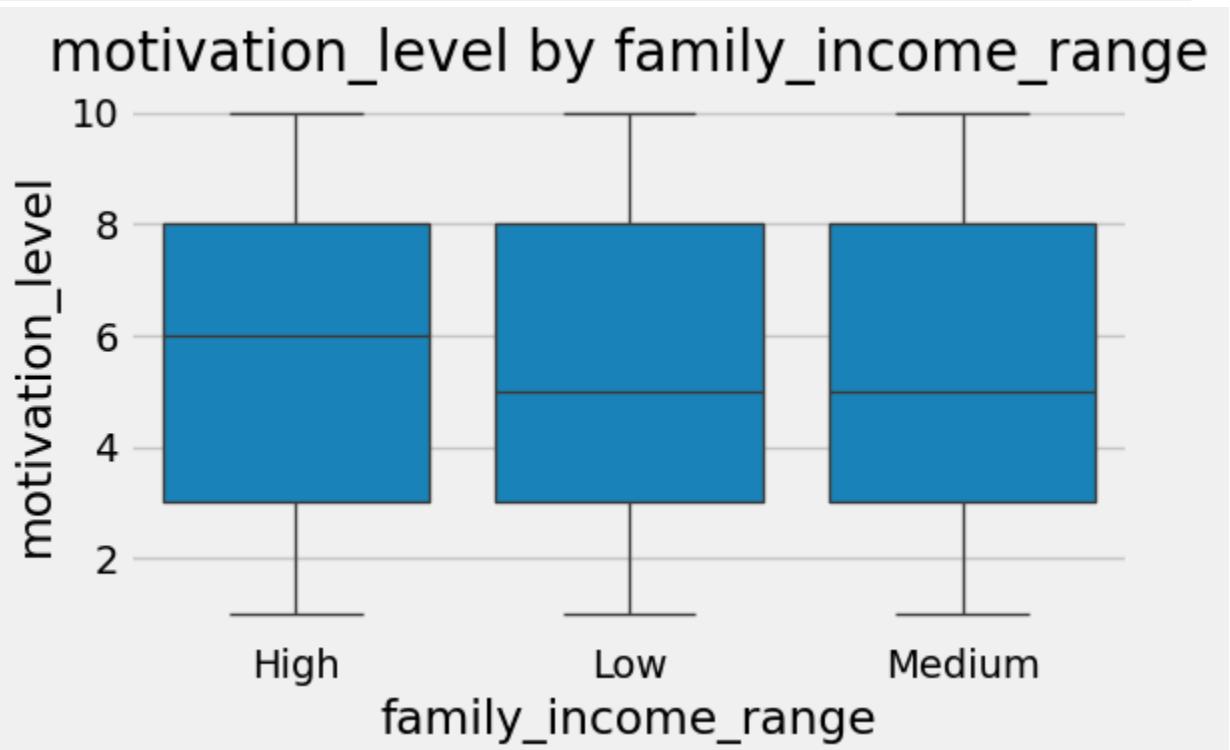
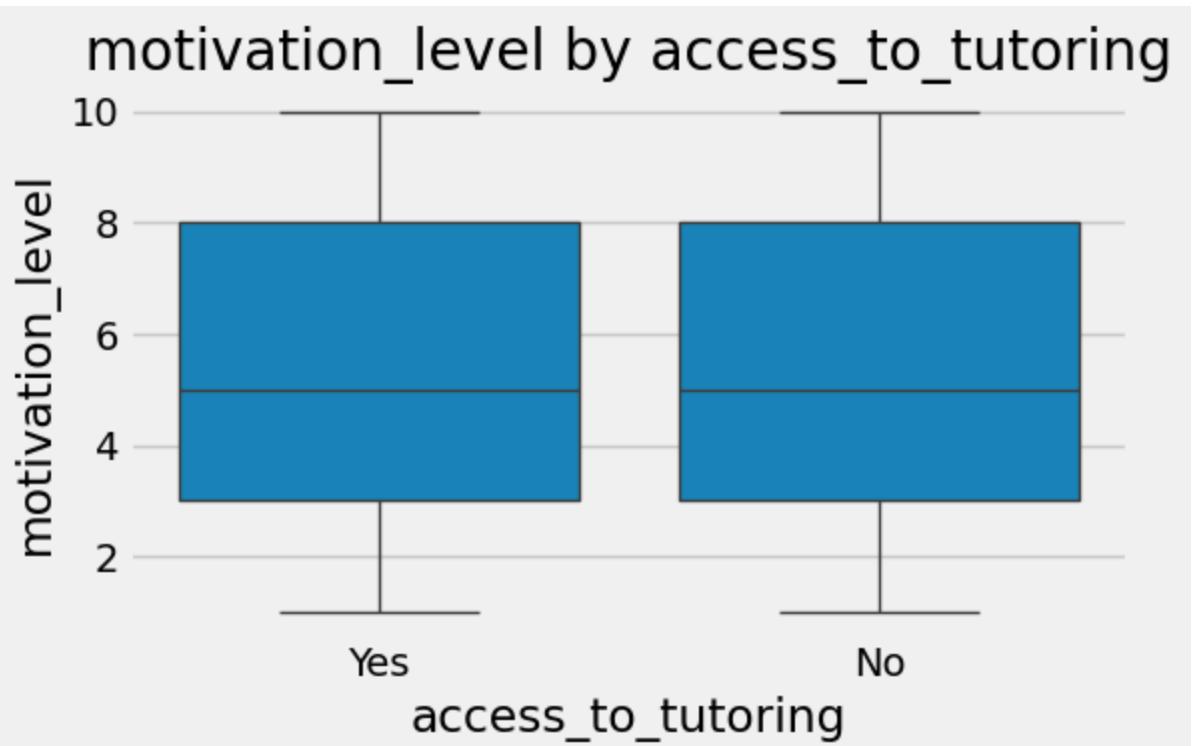


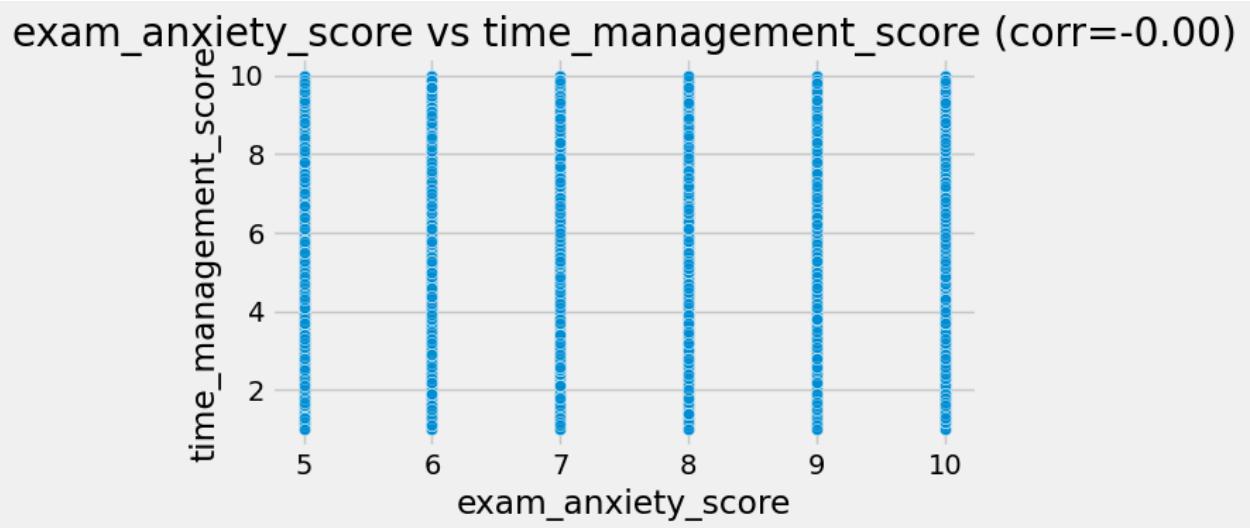
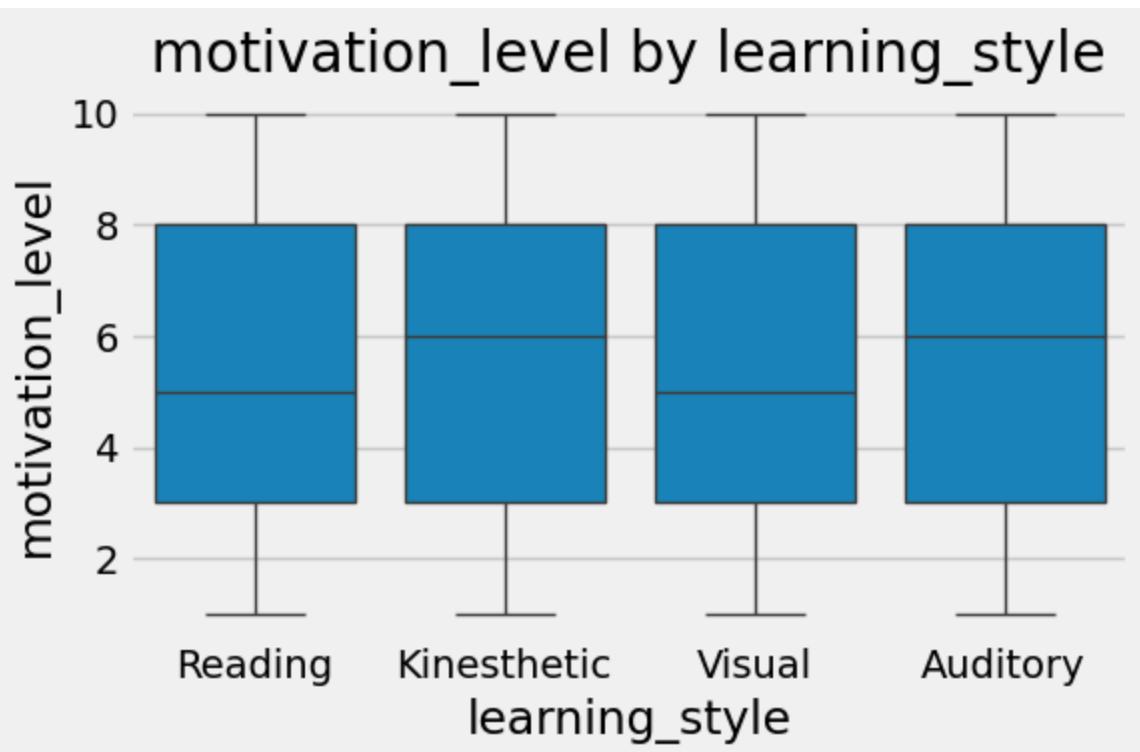


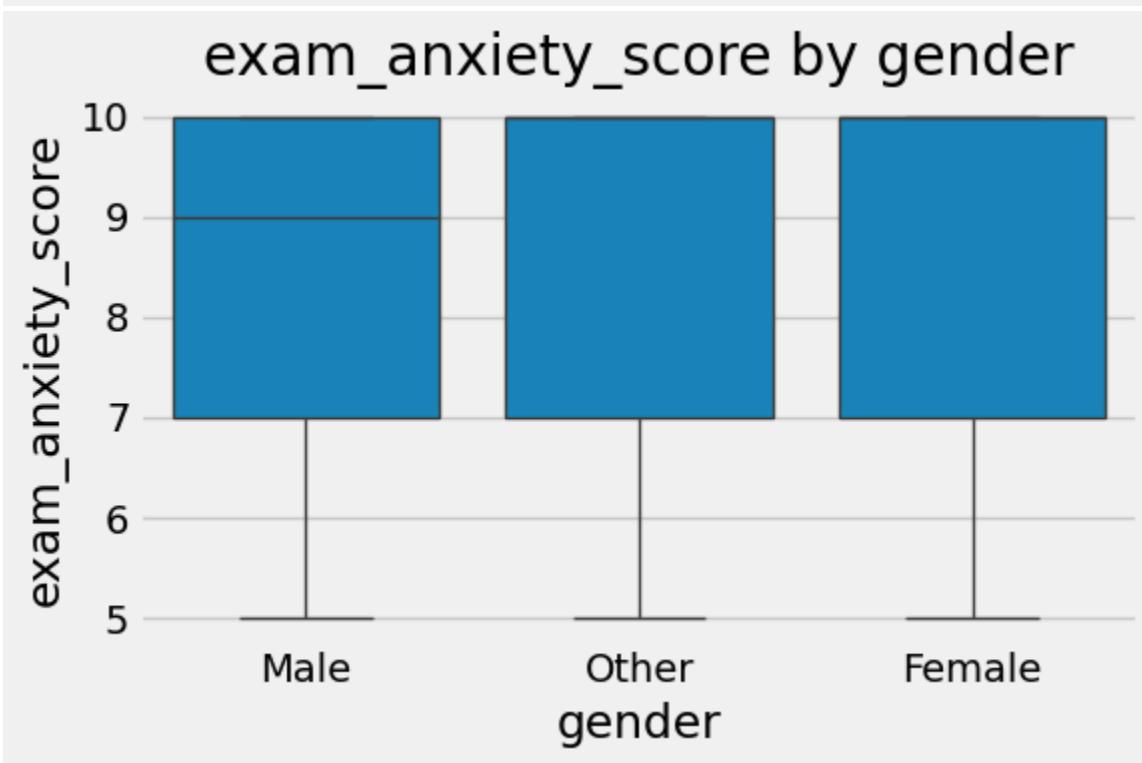
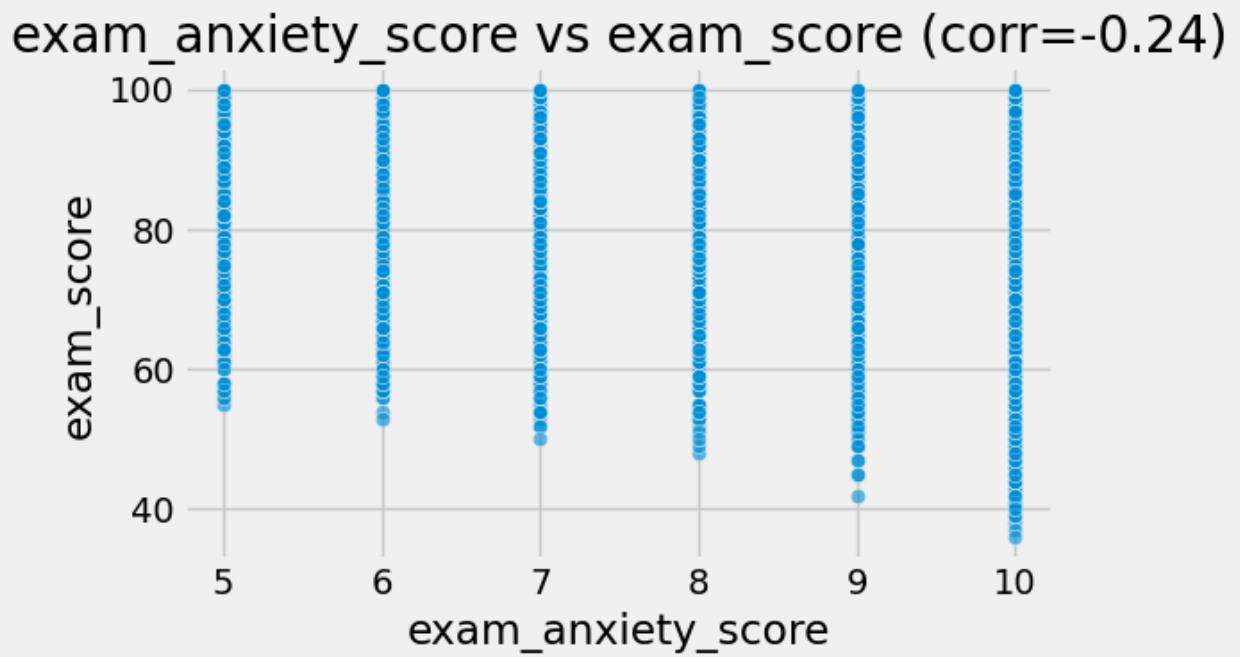


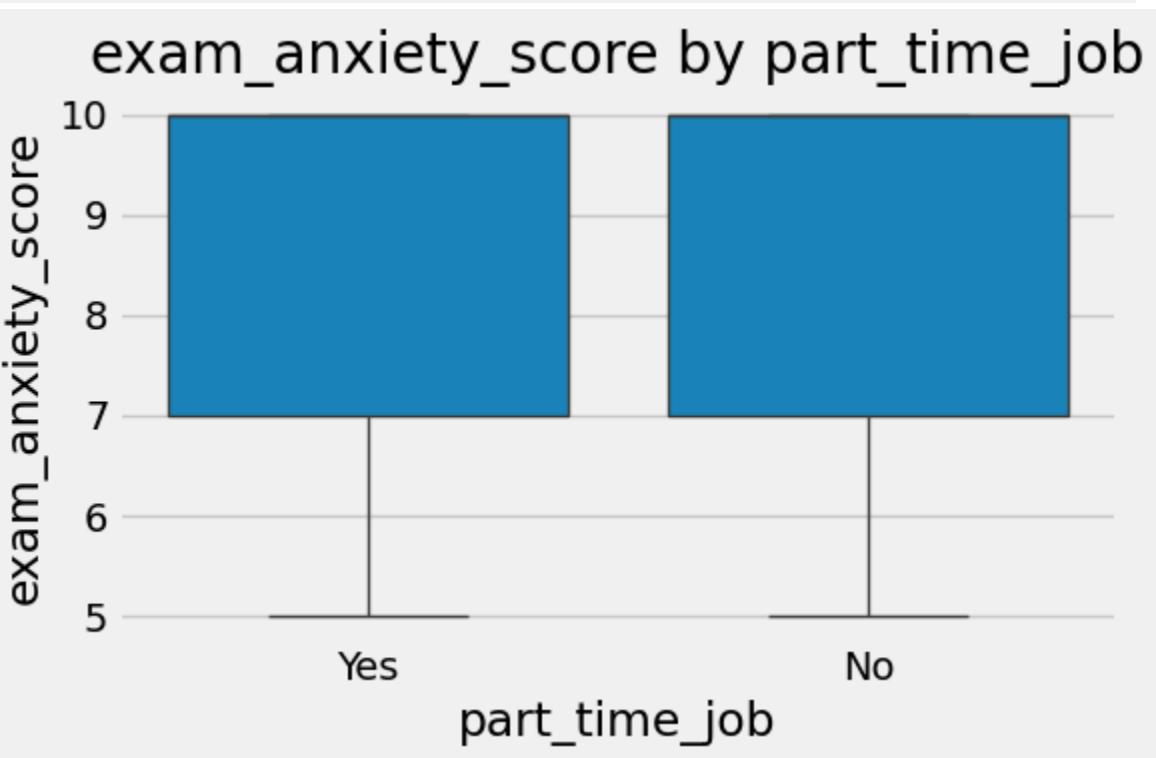
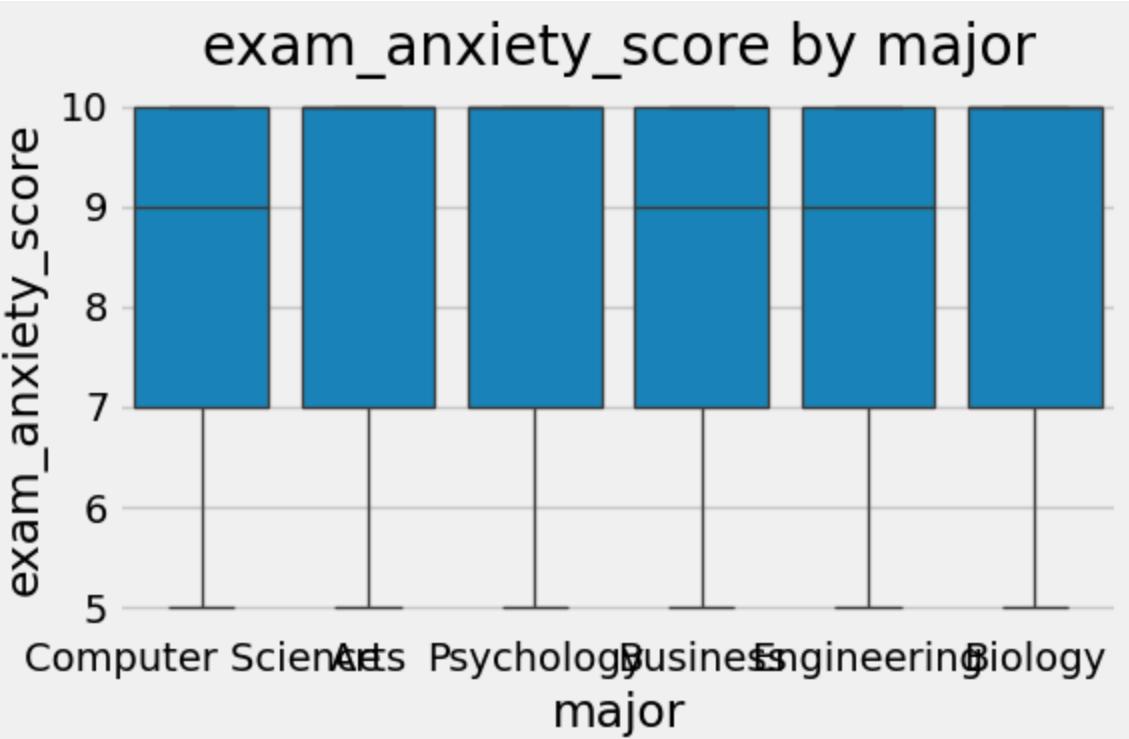


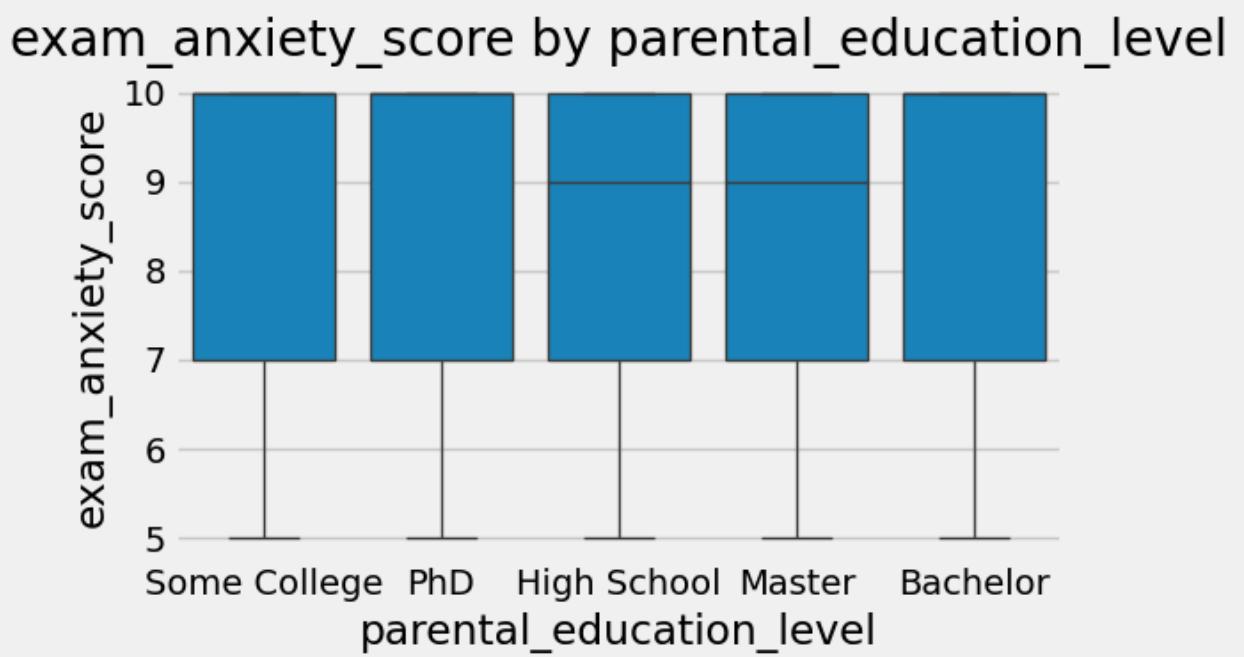
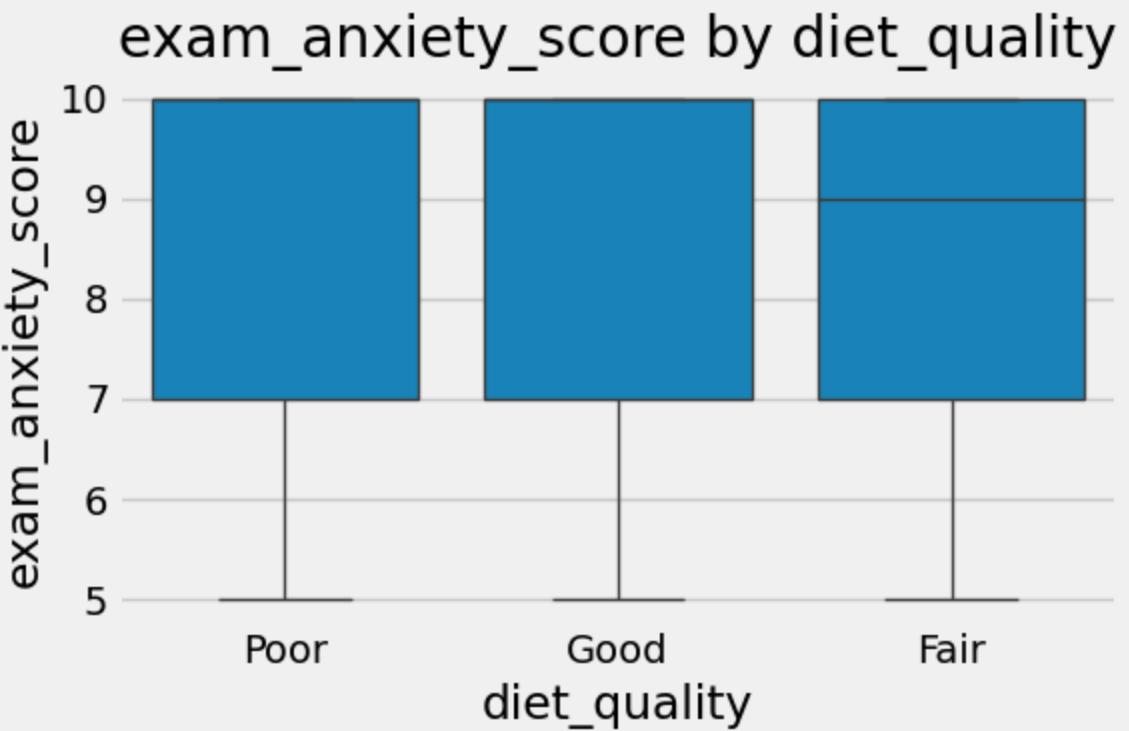


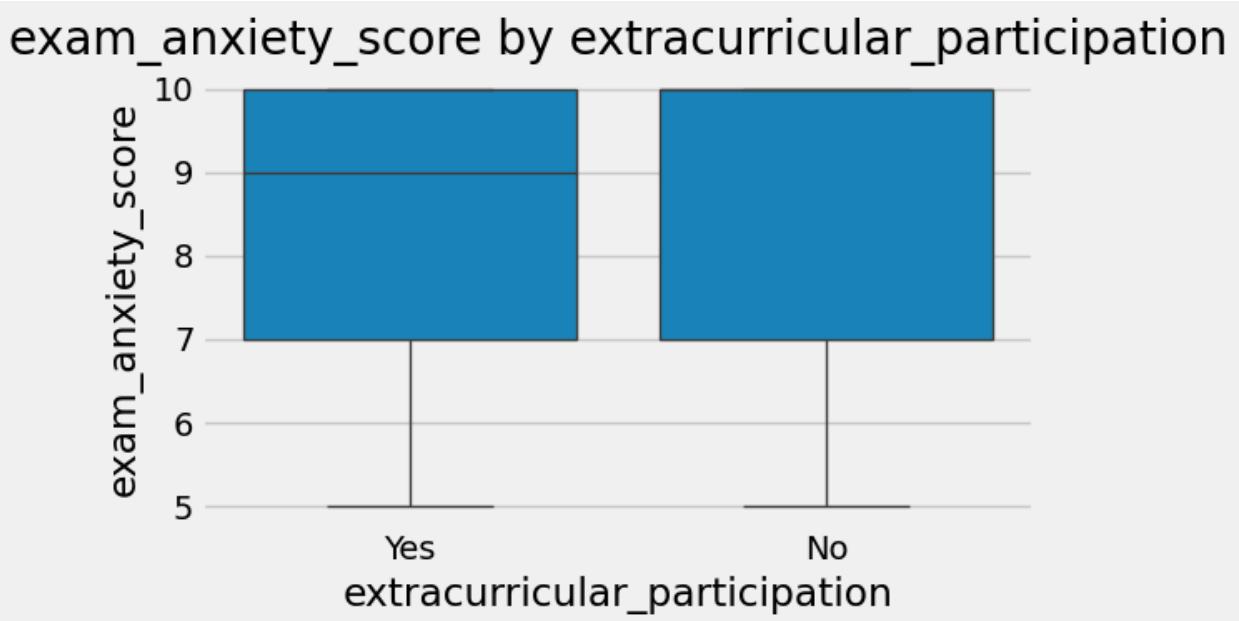
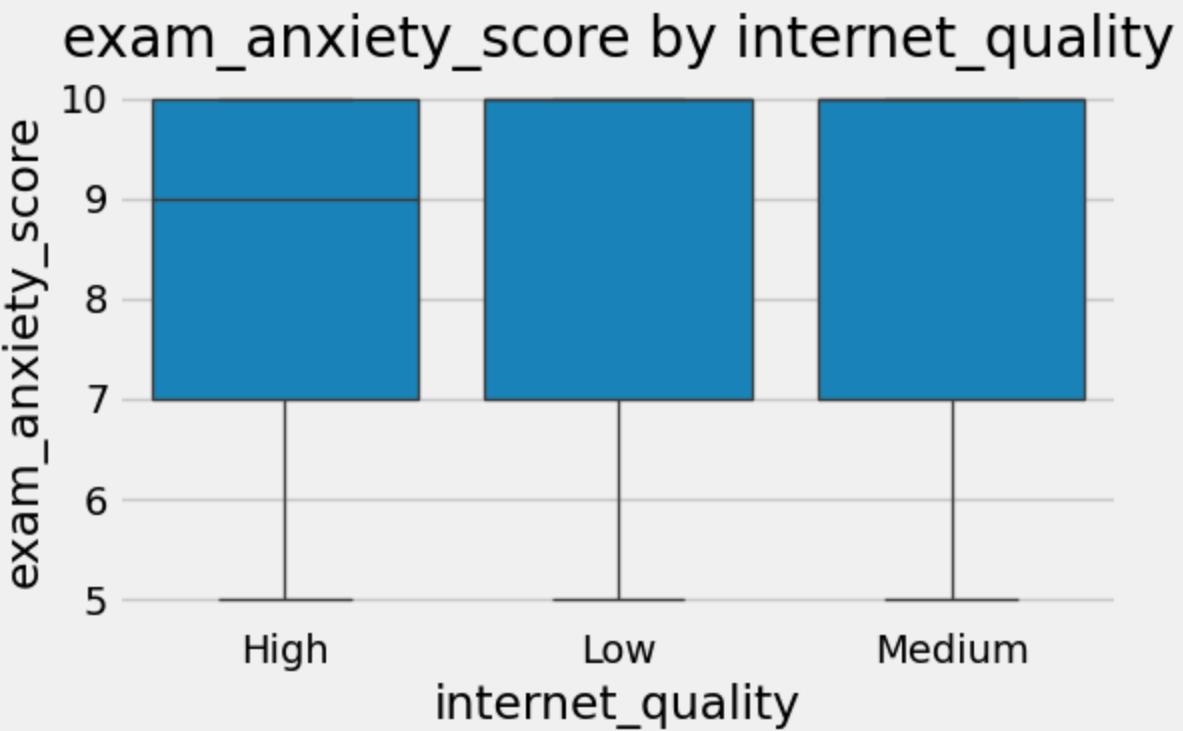


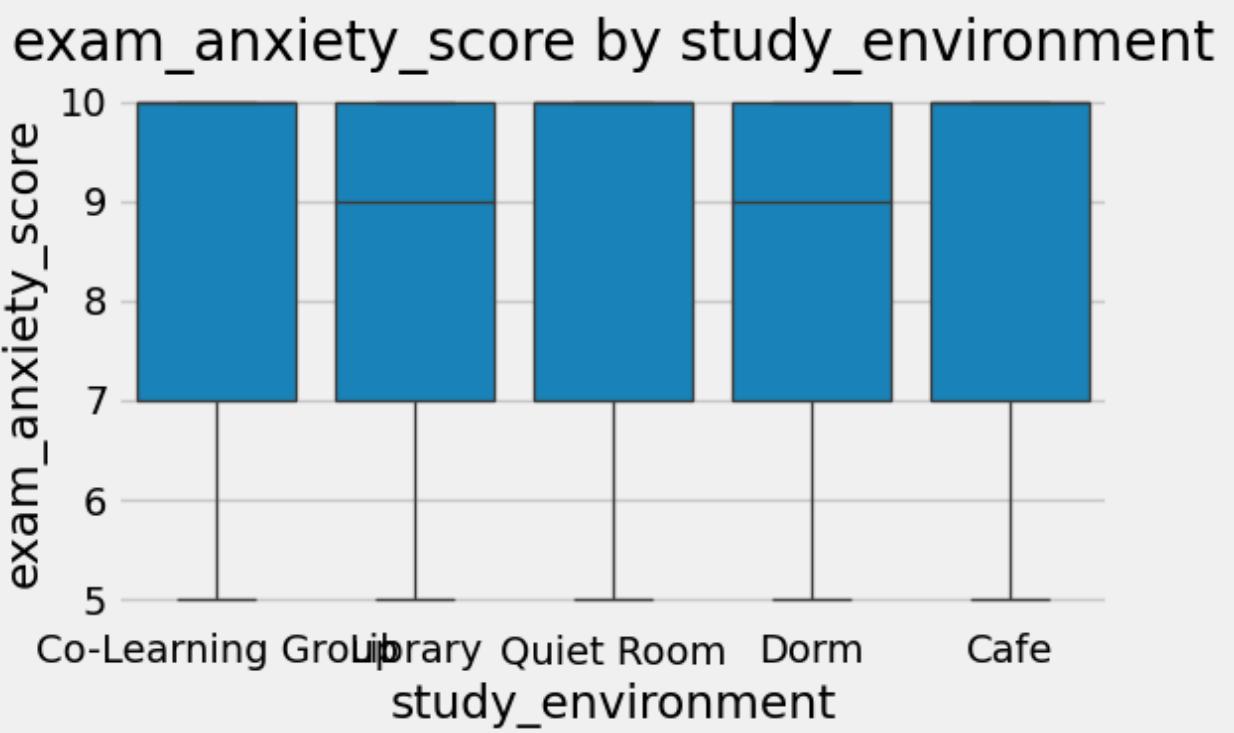
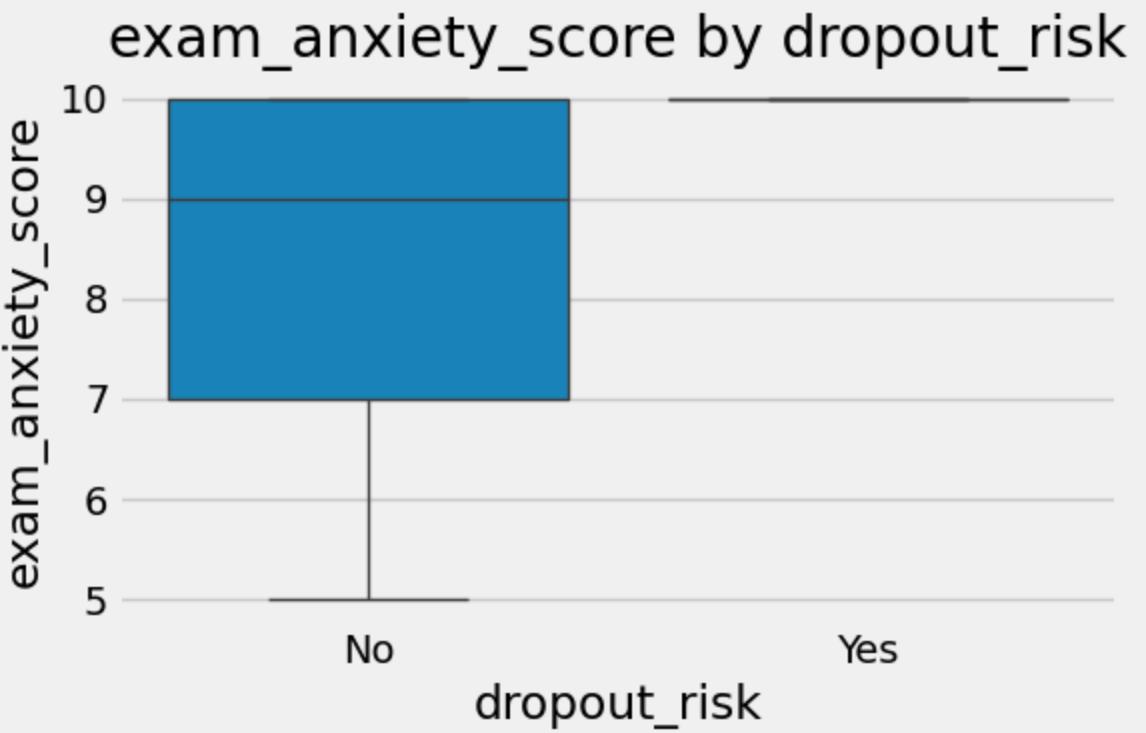




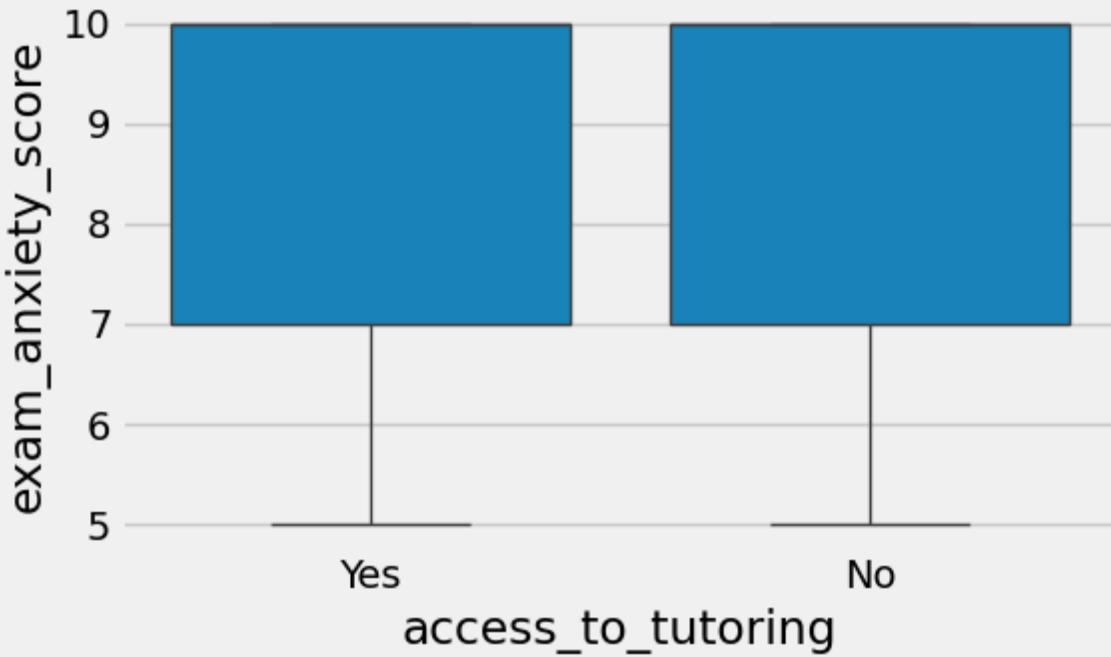




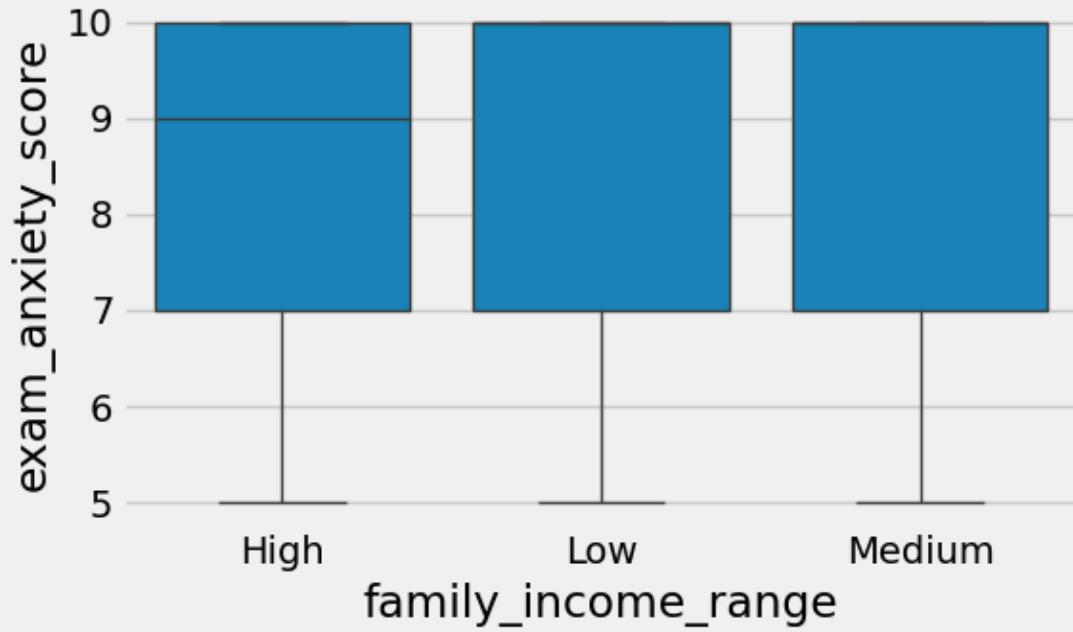


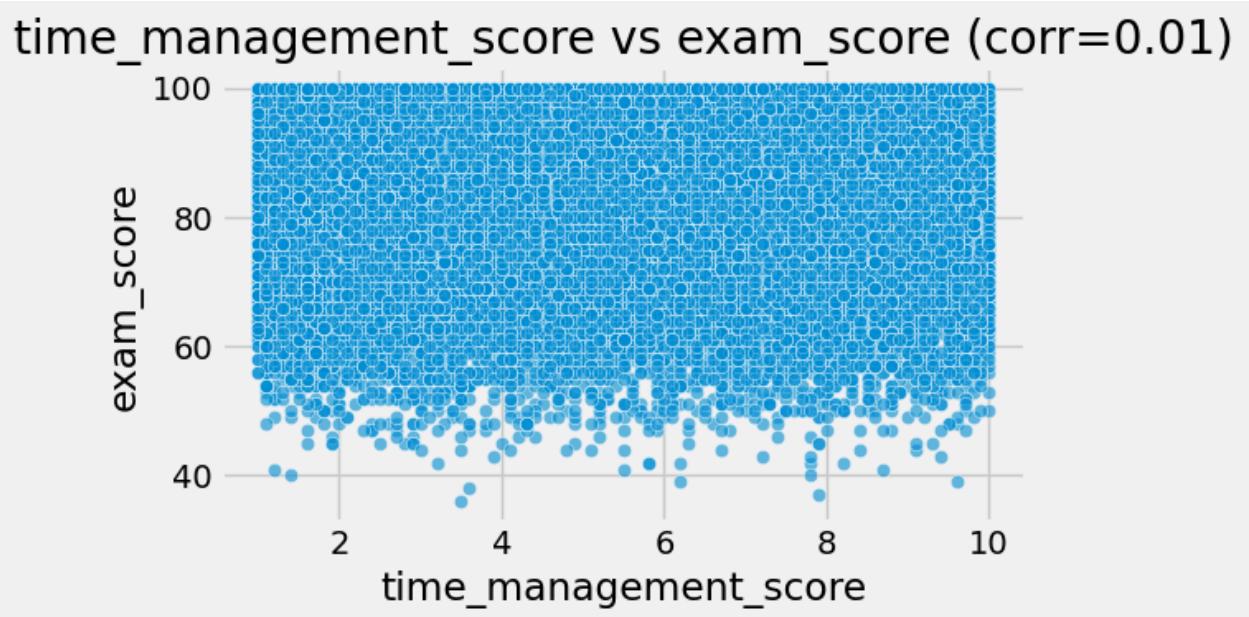
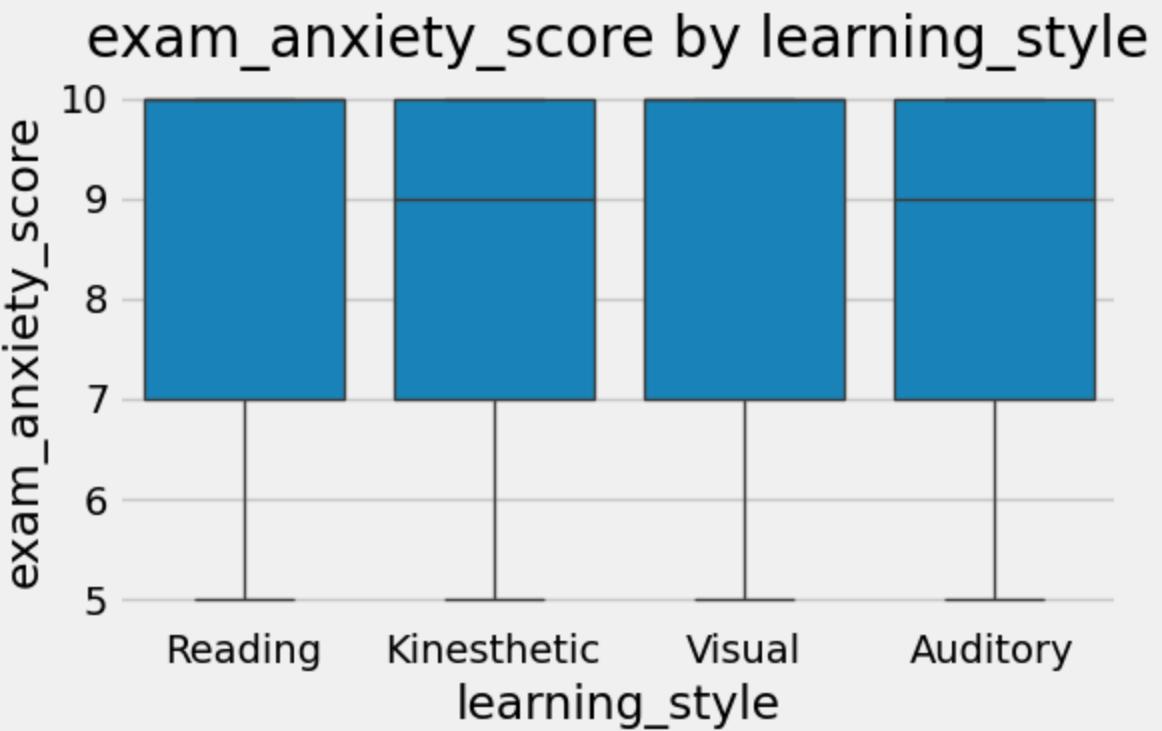


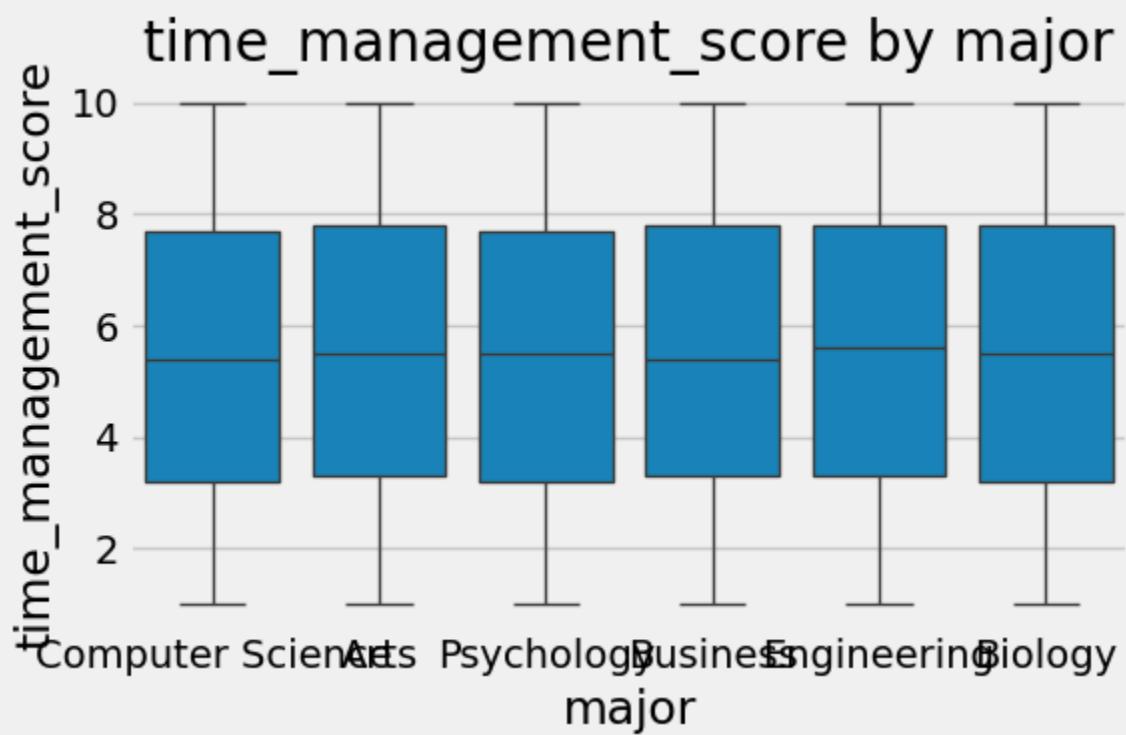
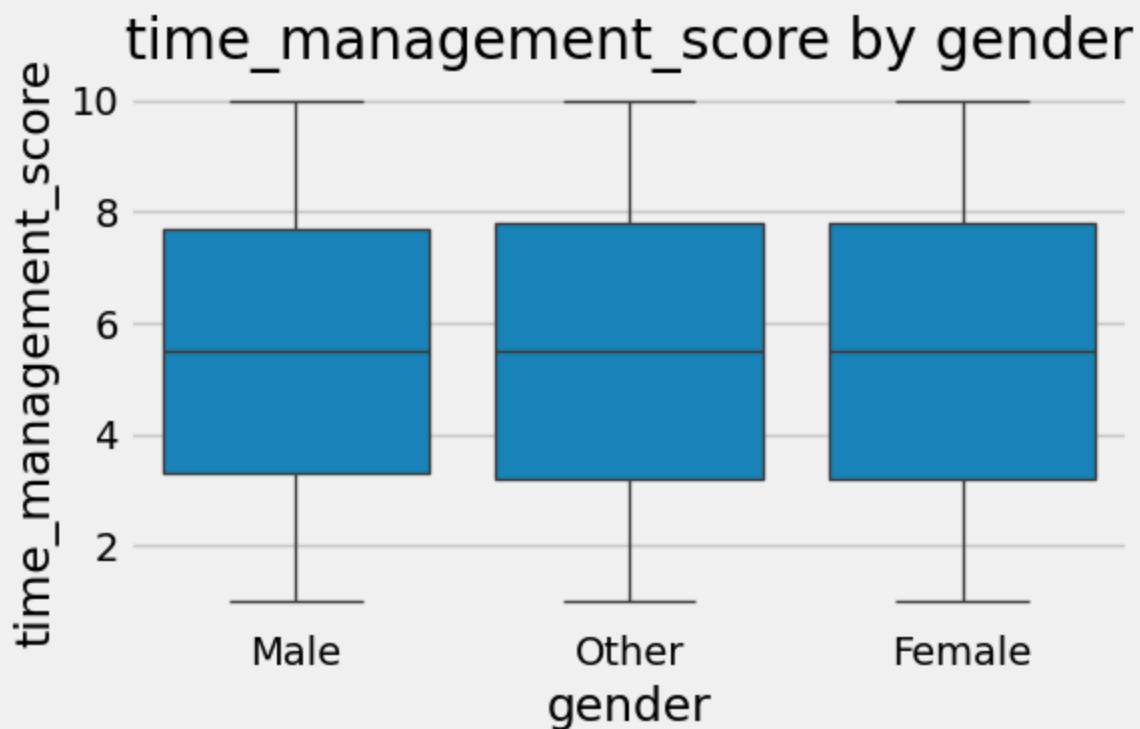
exam_anxiety_score by access_to_tutoring

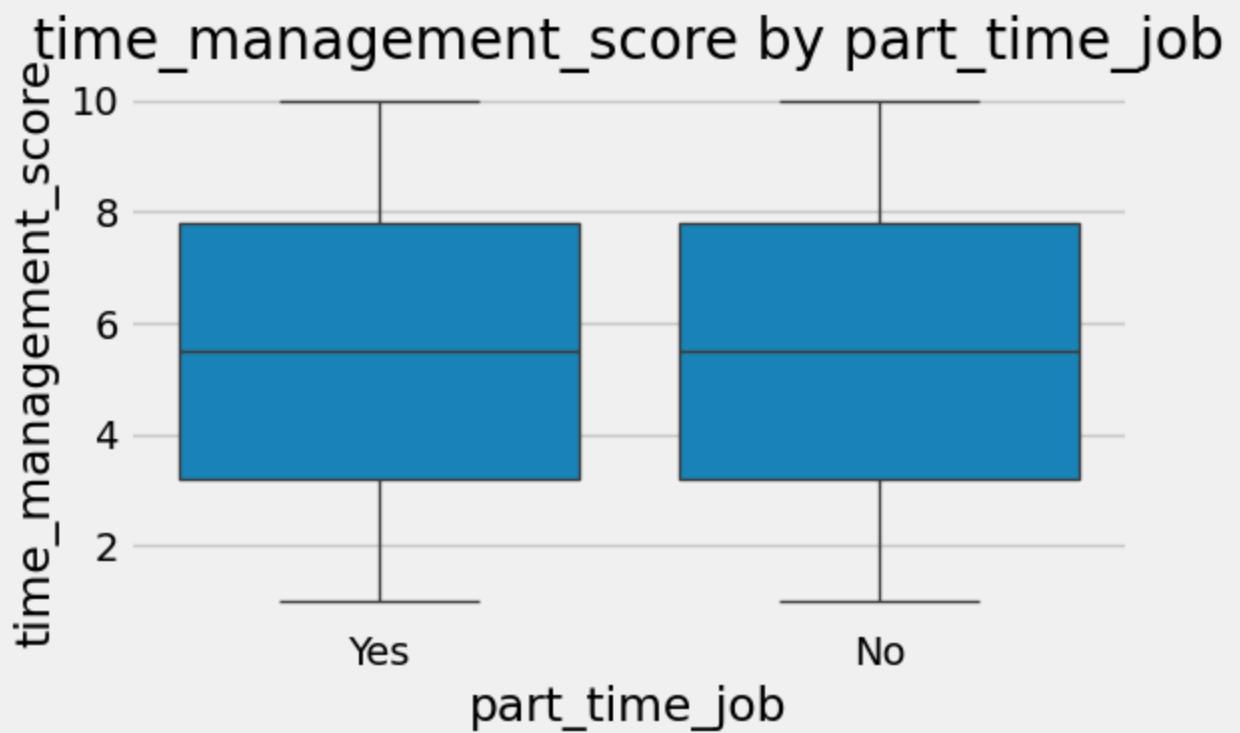


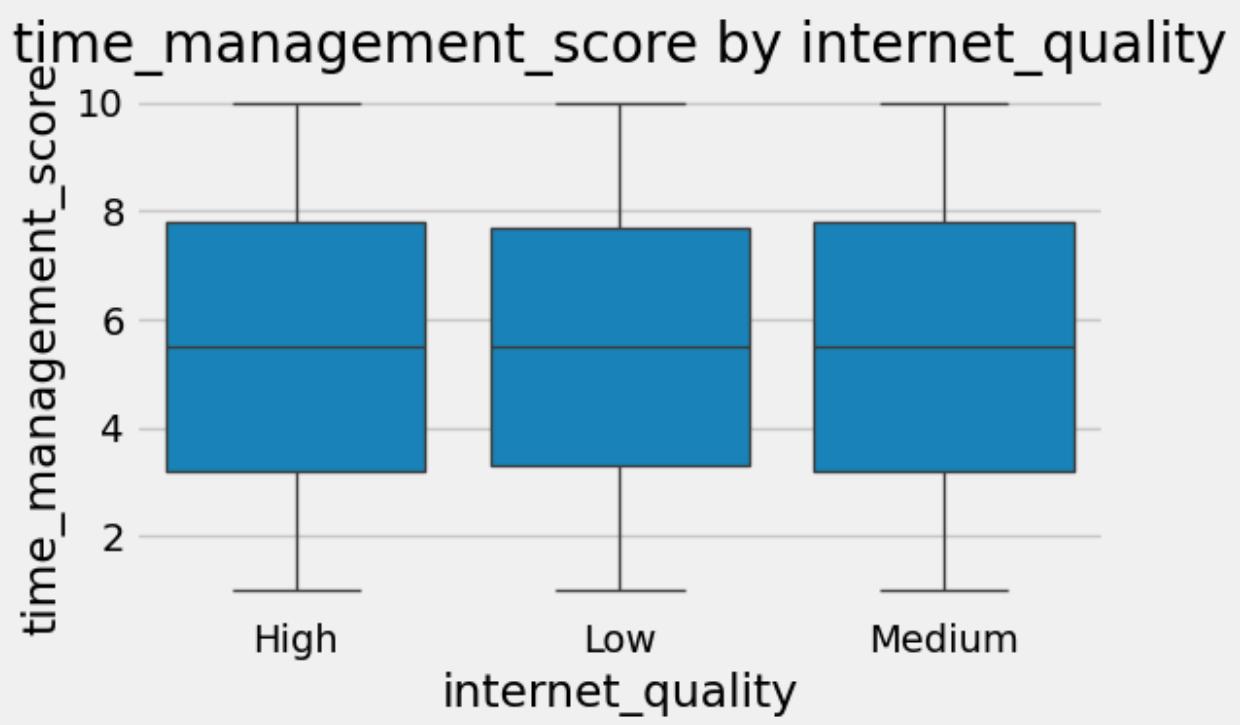
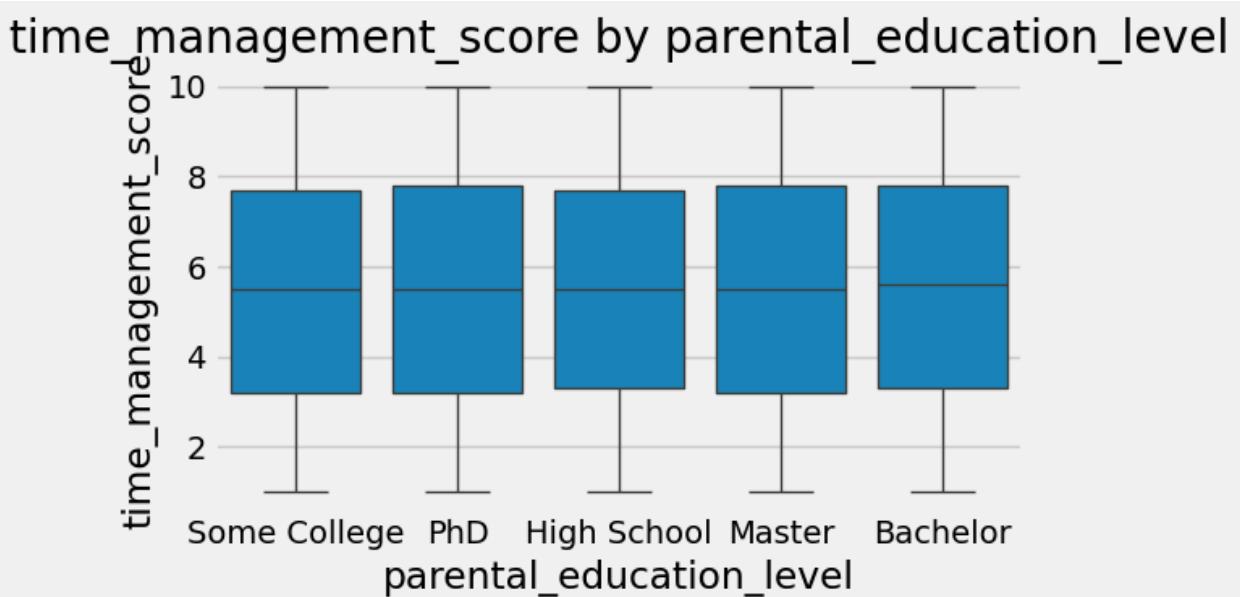
exam_anxiety_score by family_income_range

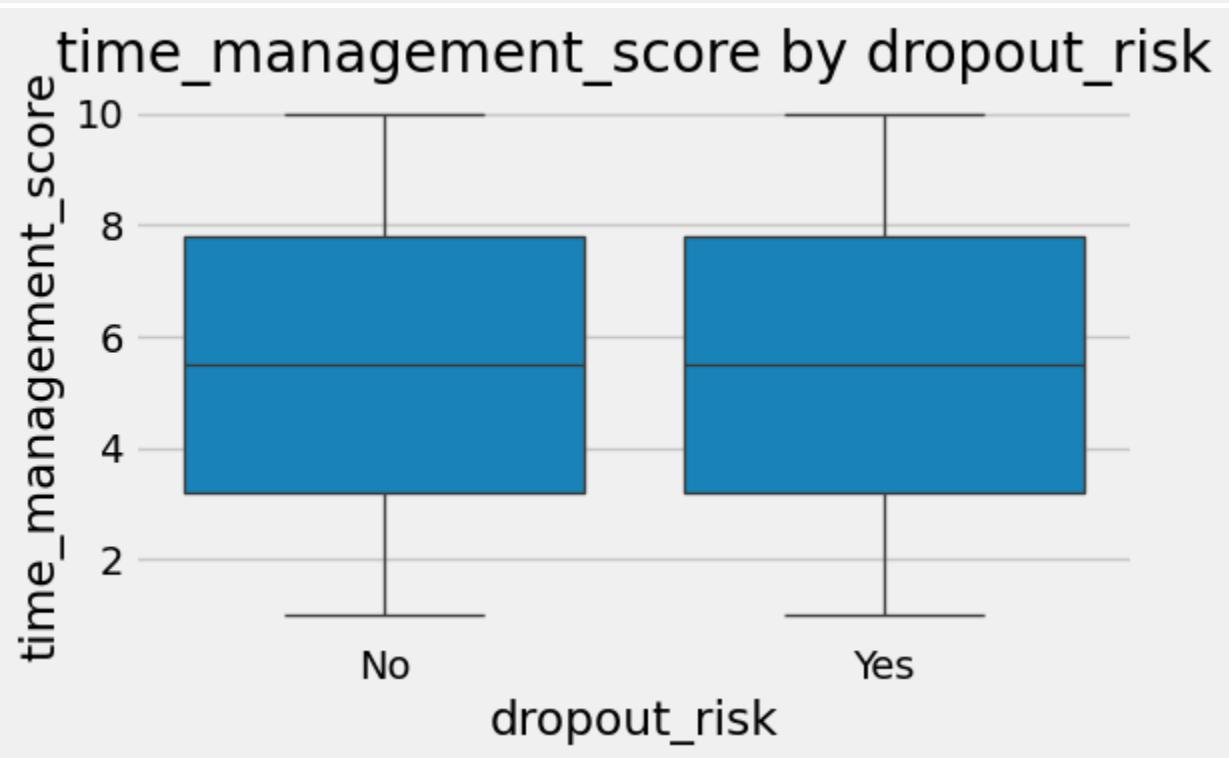
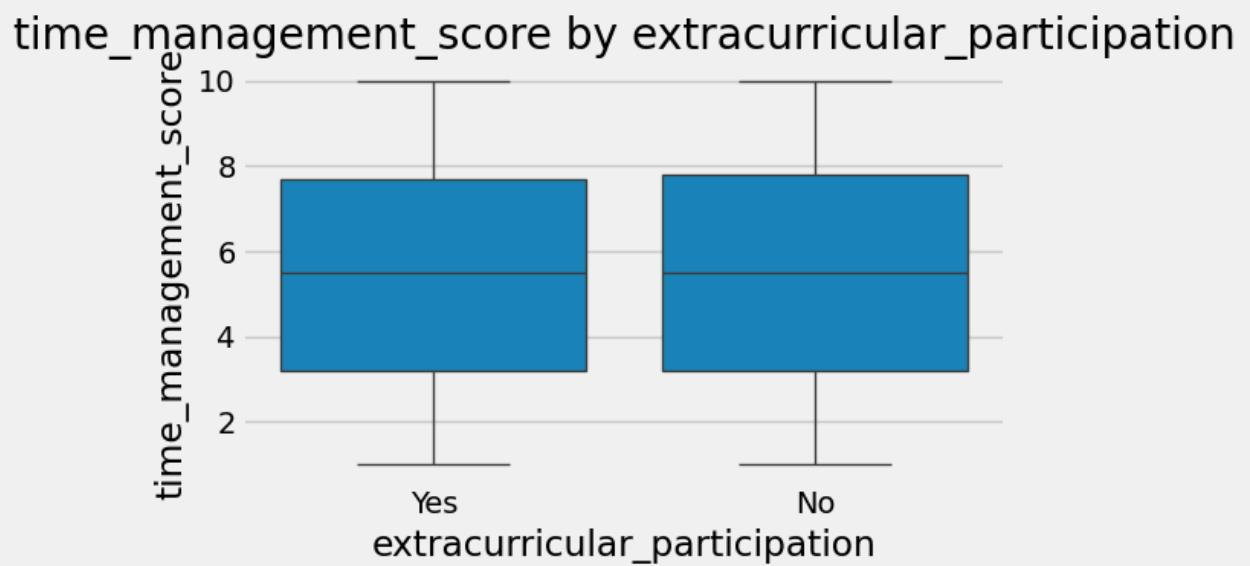


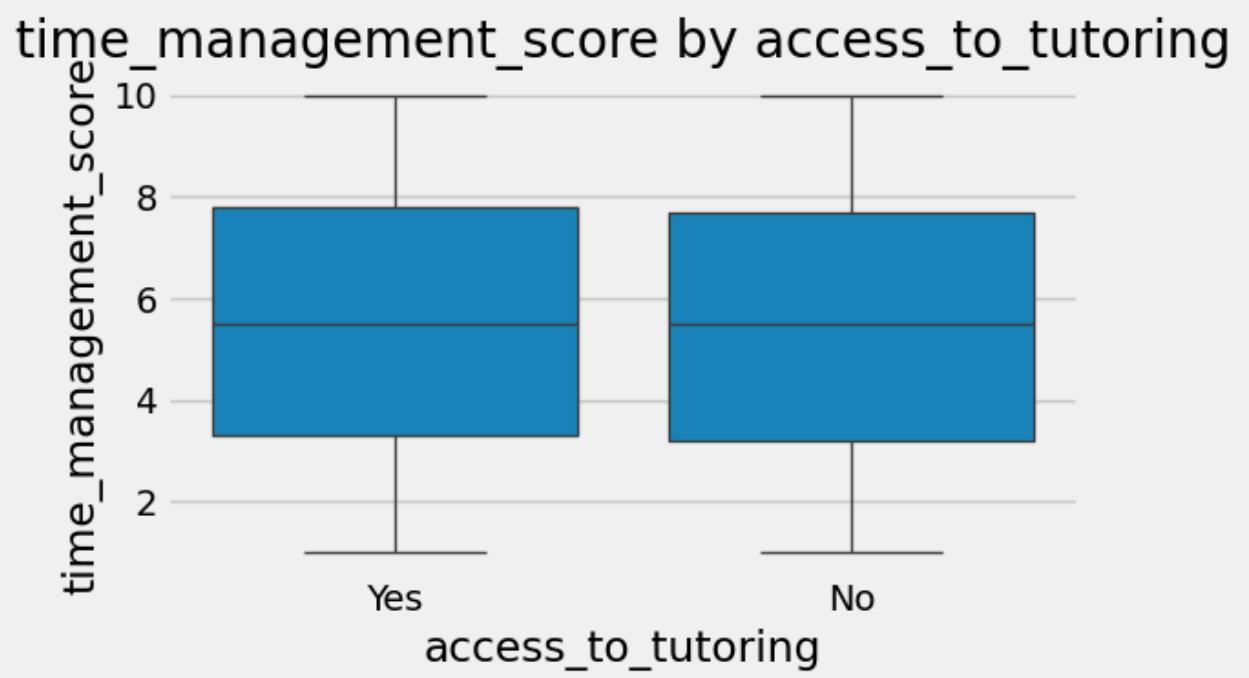
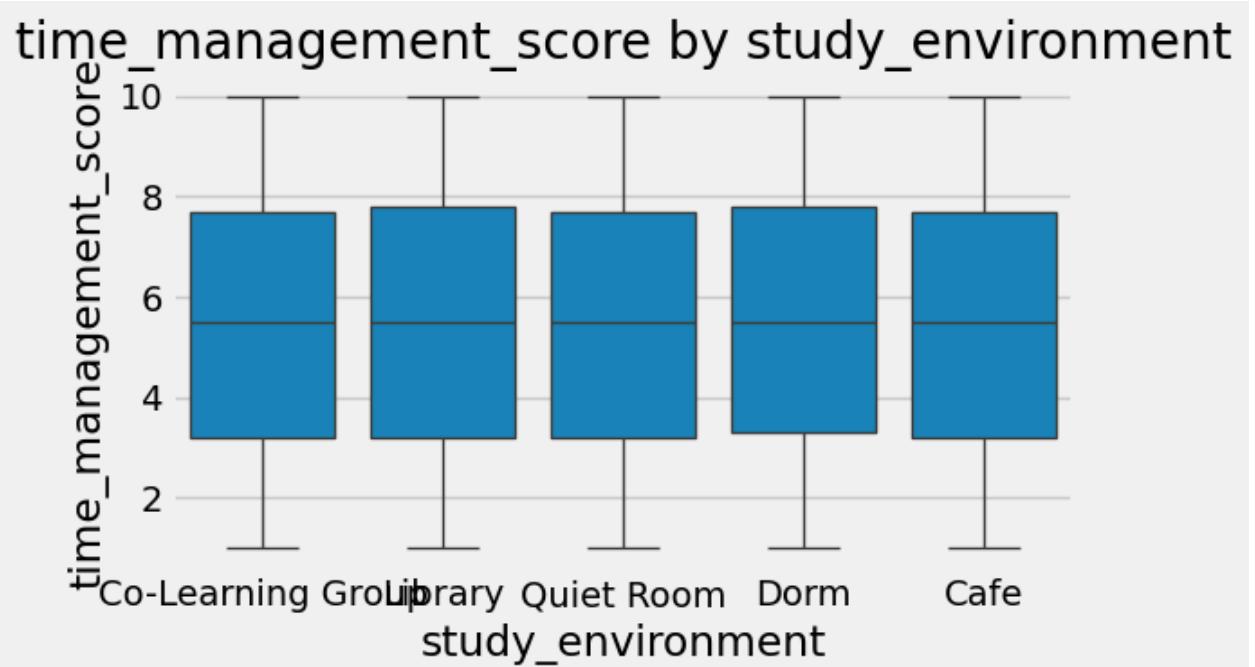


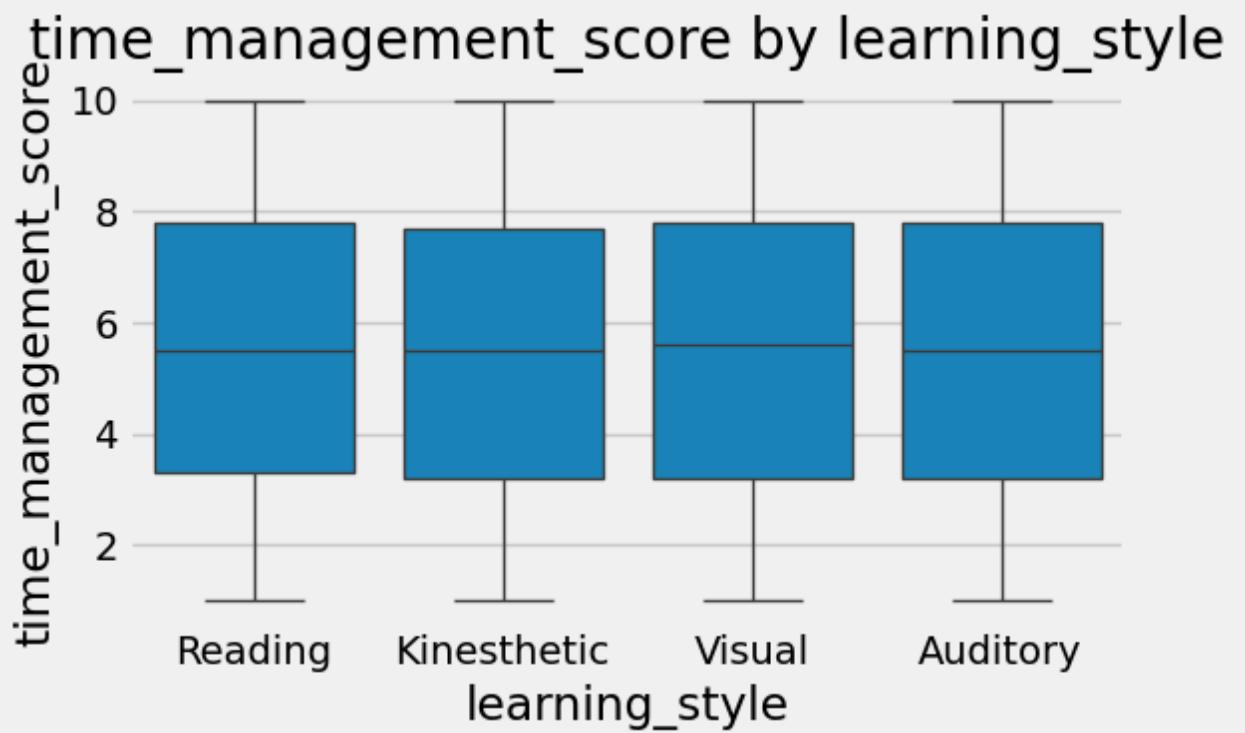
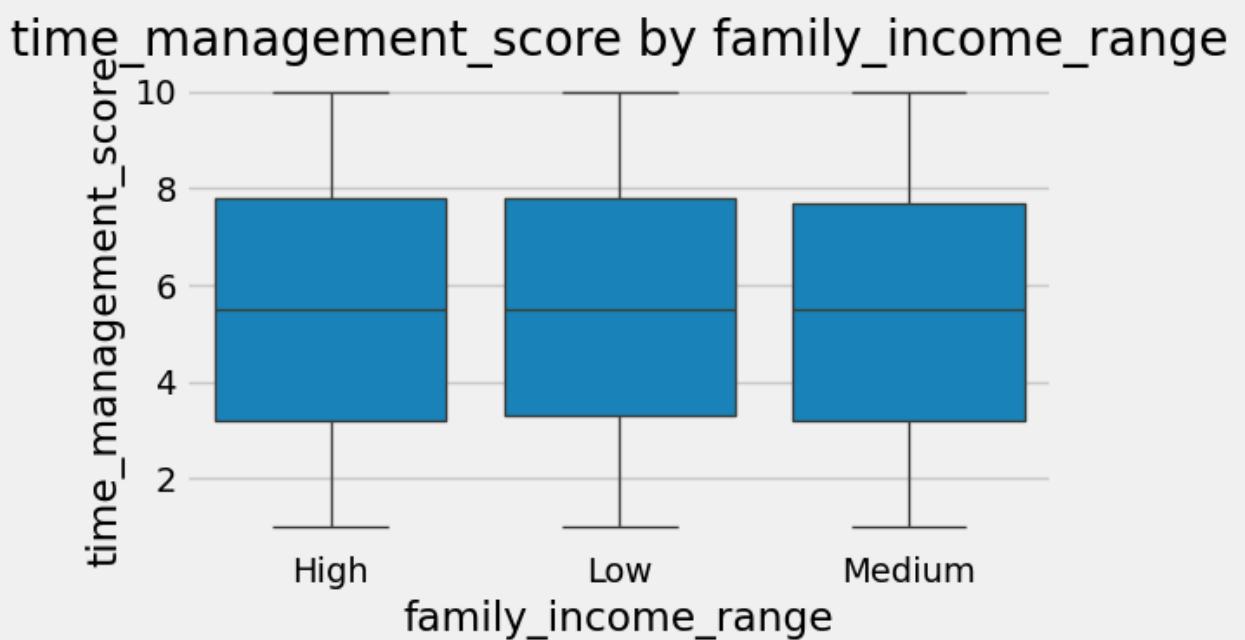


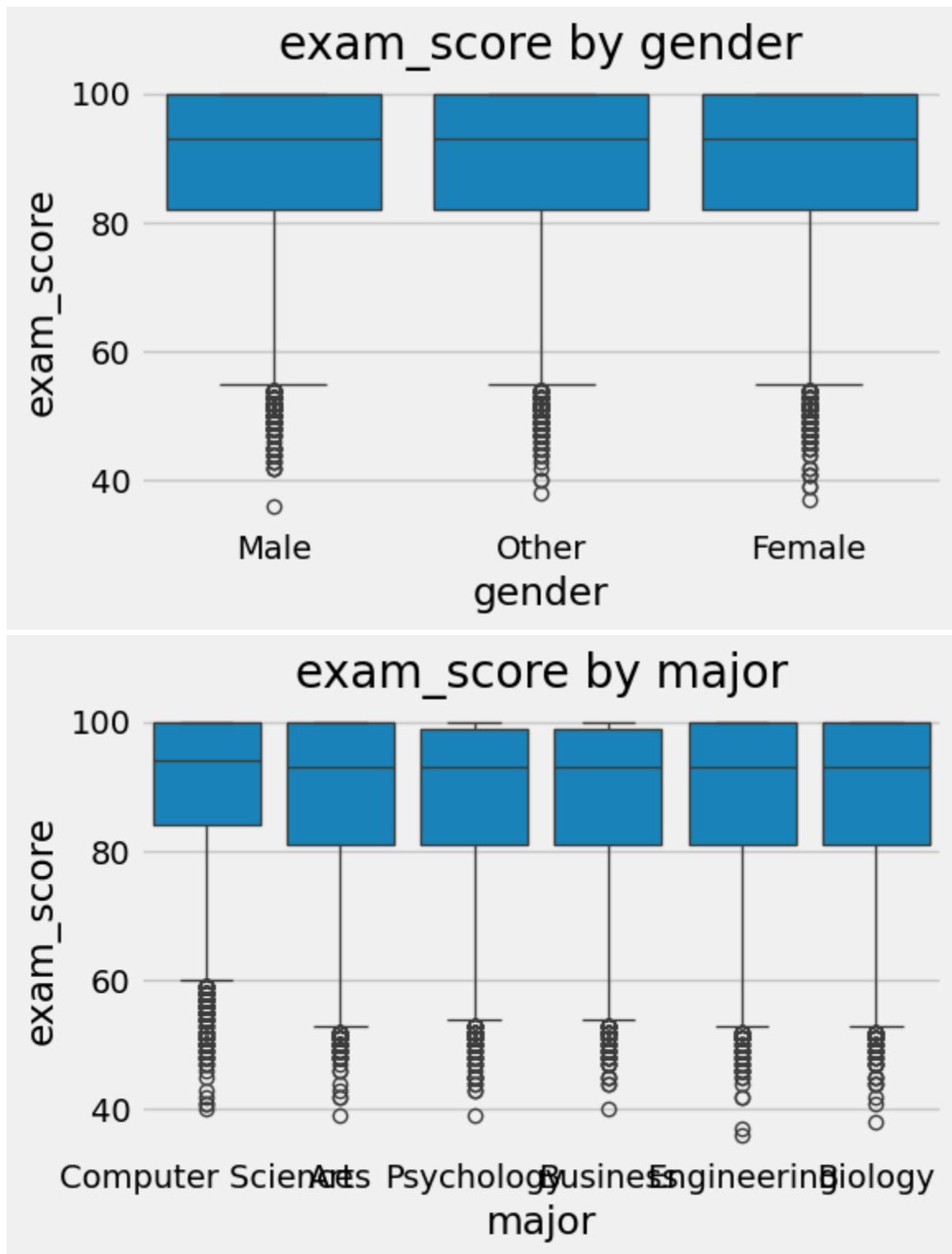


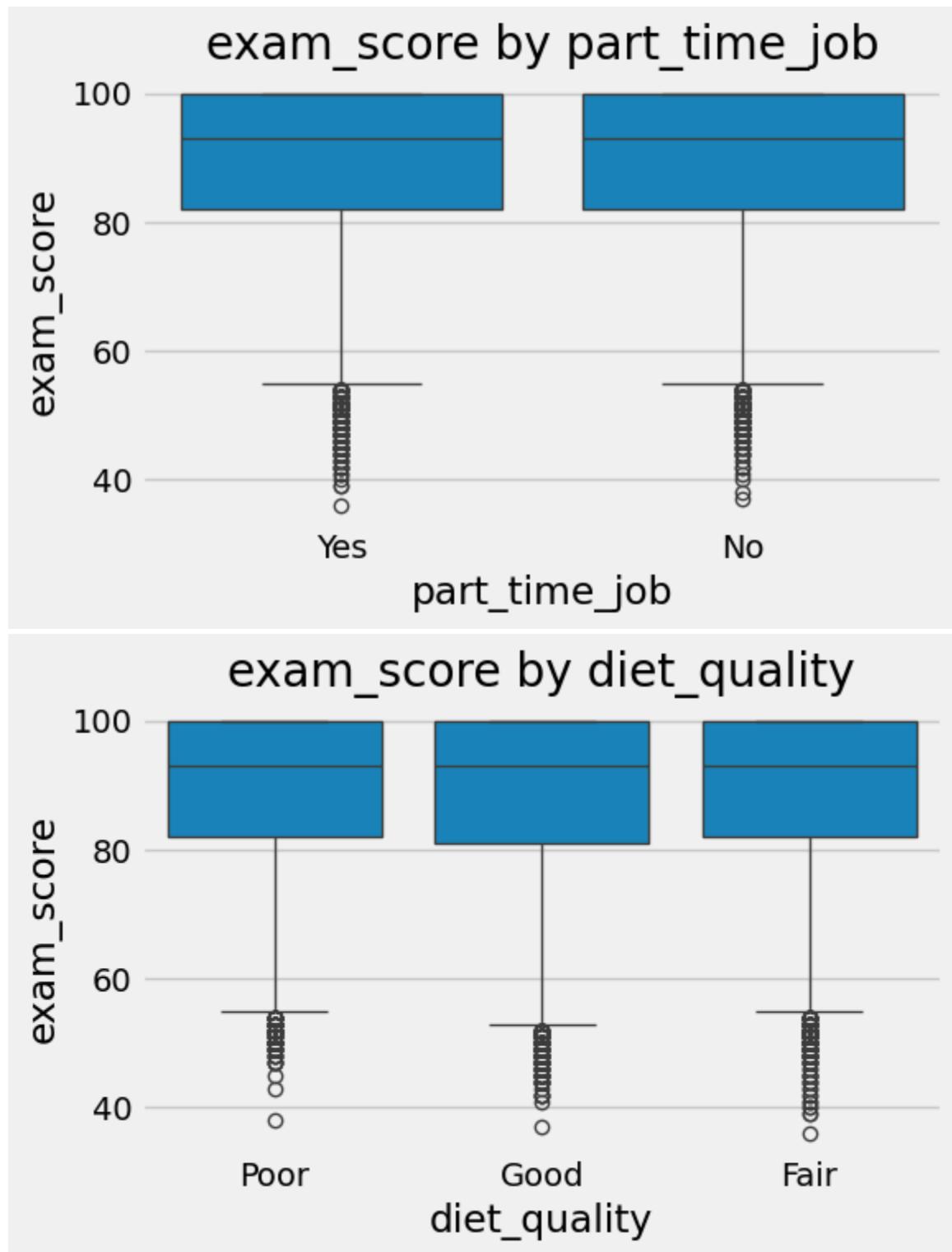




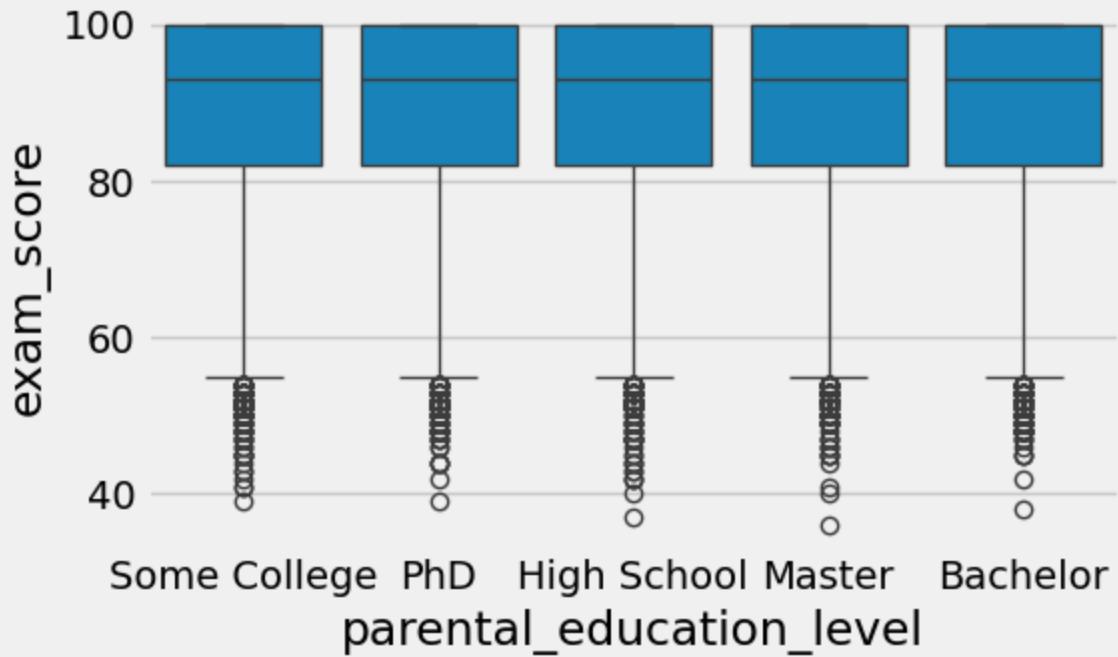




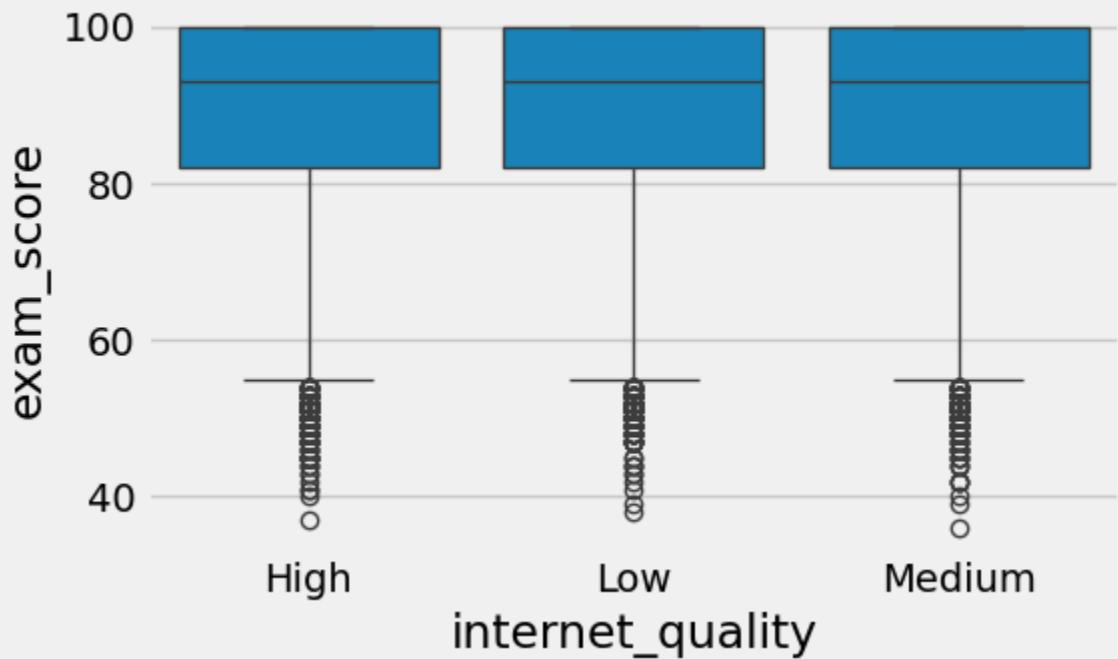


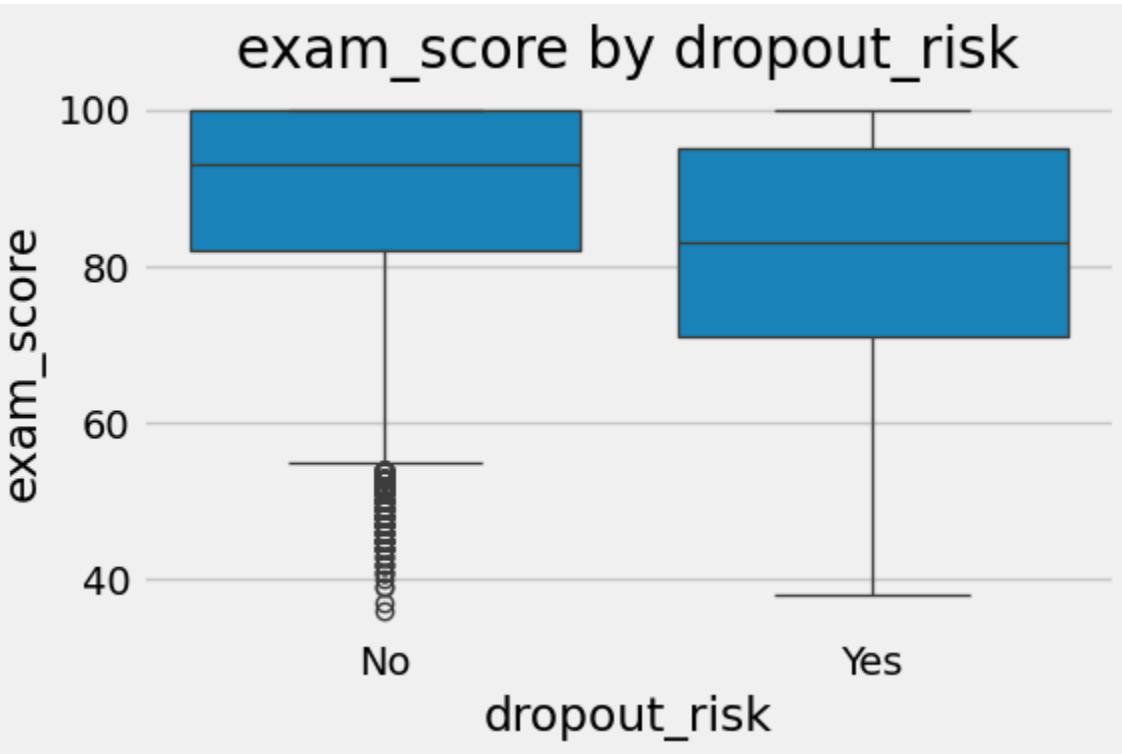
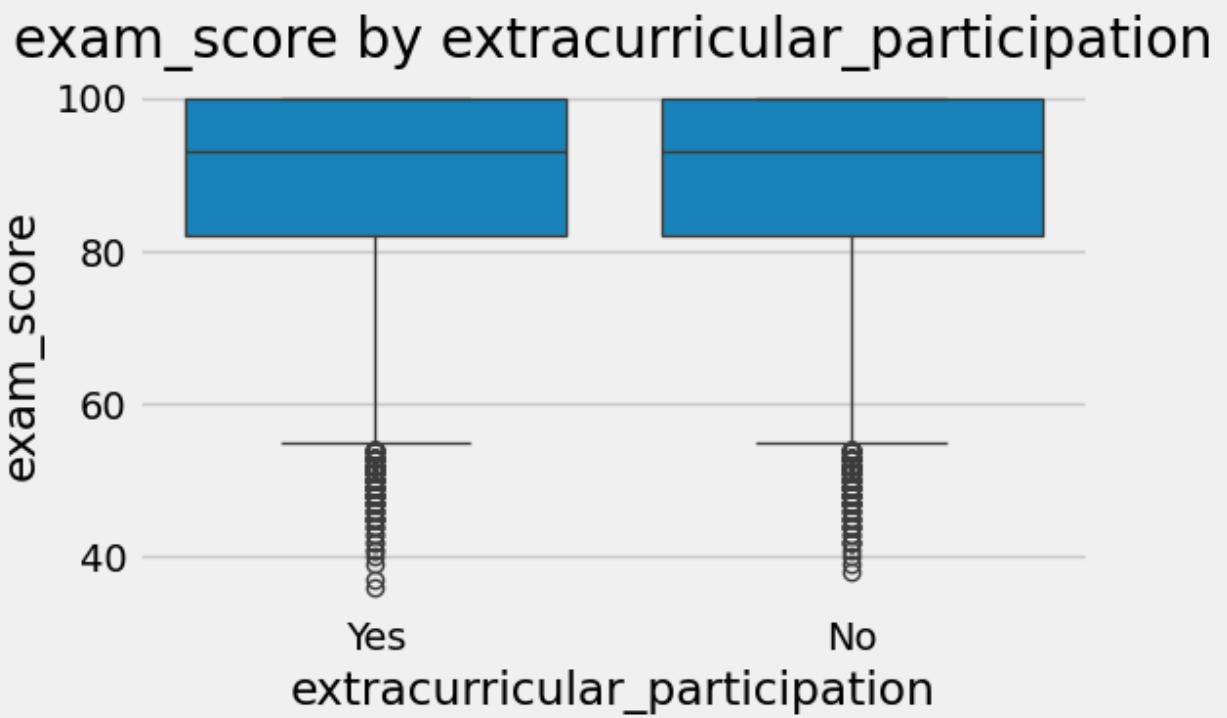


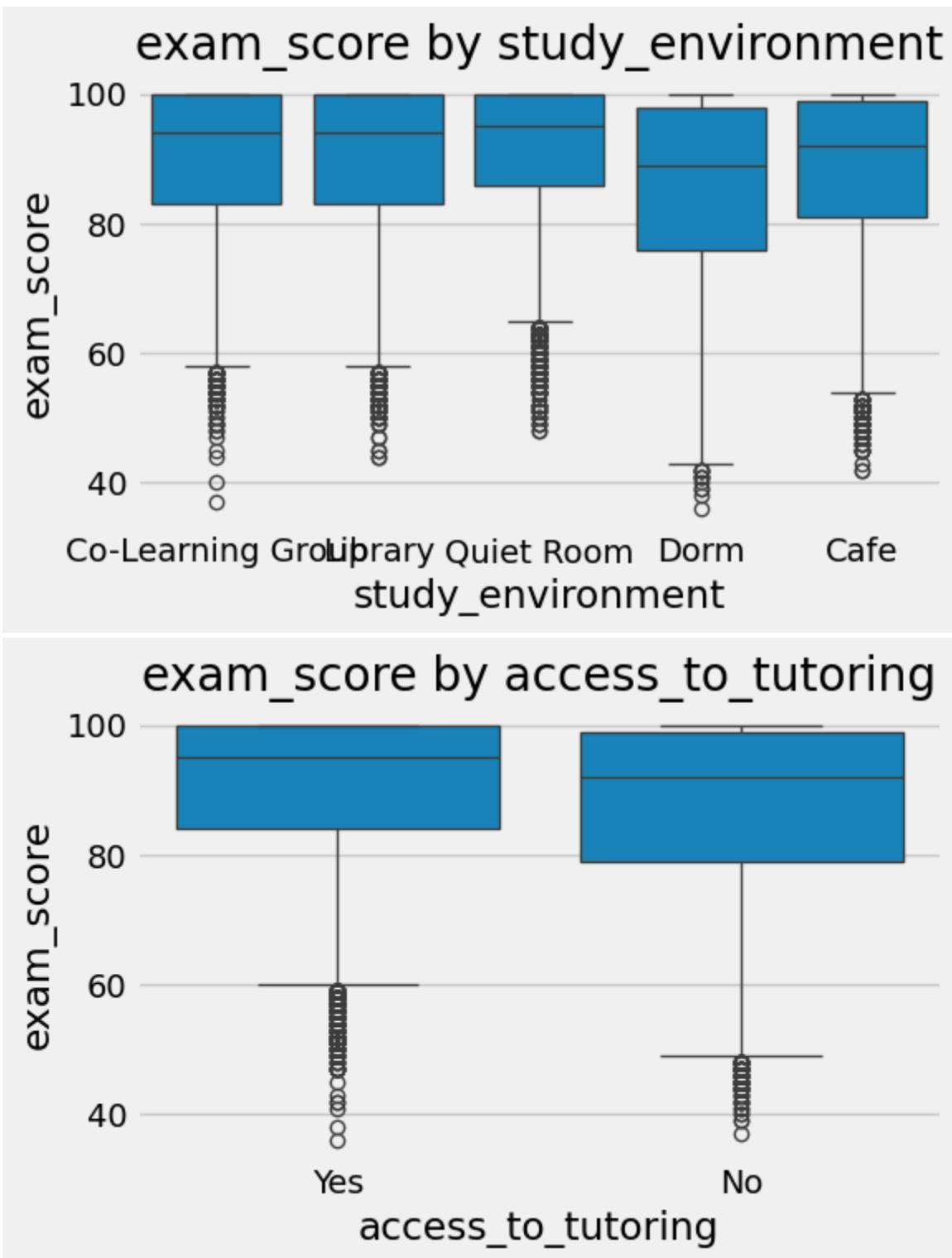
exam_score by parental_education_level

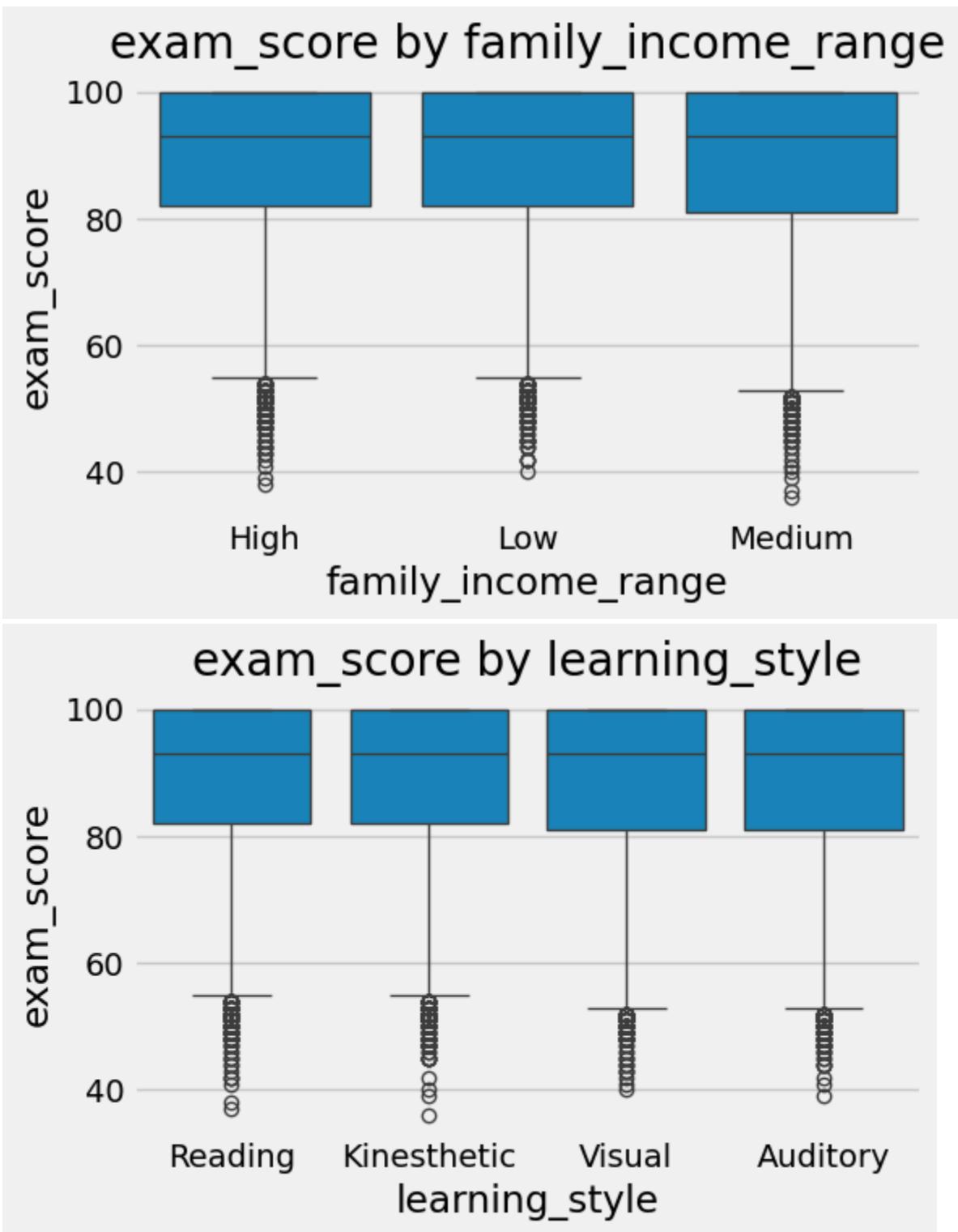


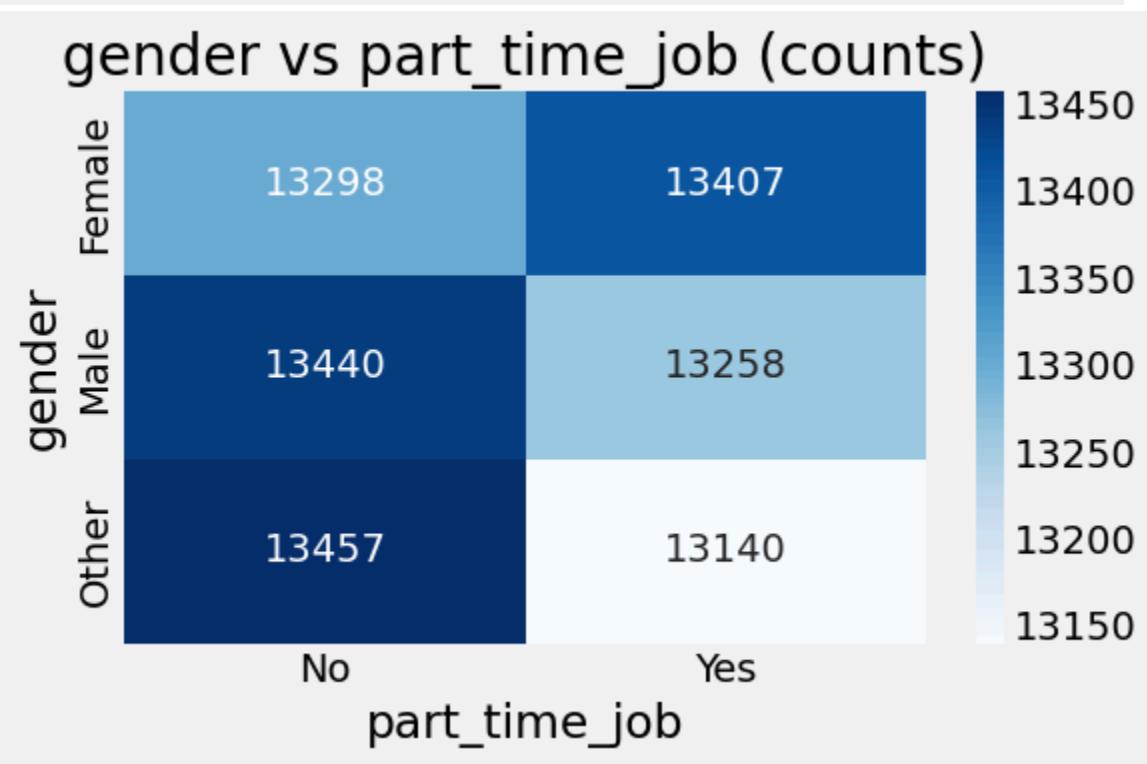
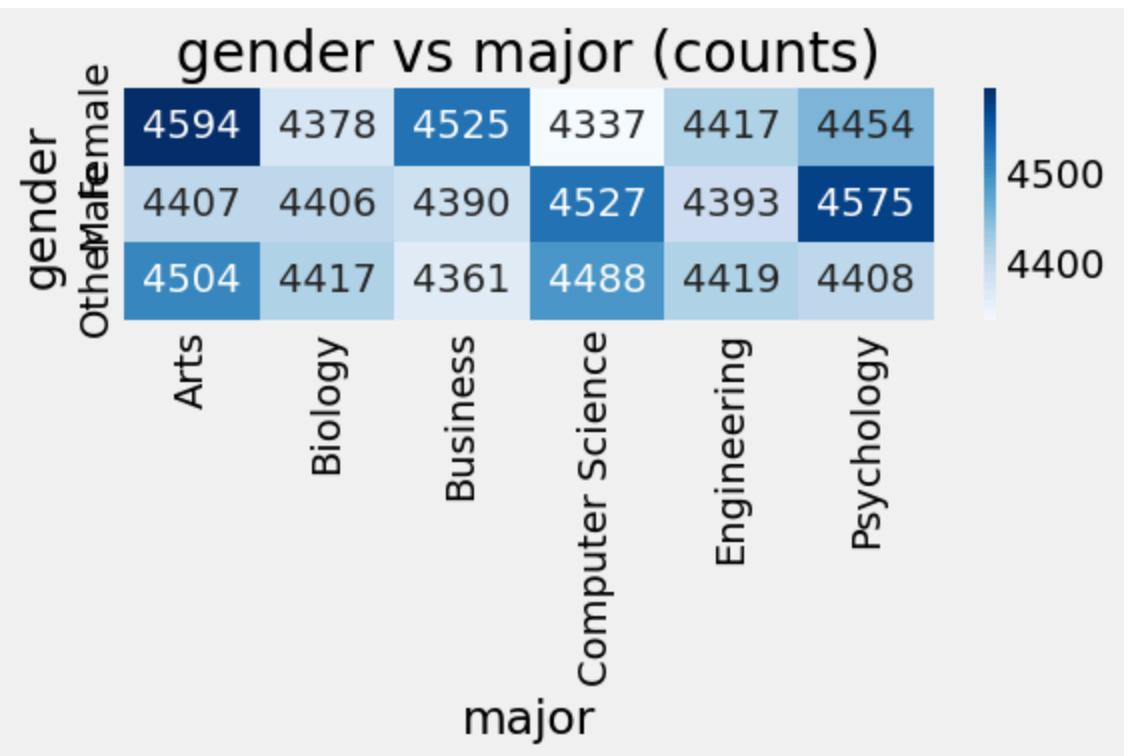
exam_score by internet_quality

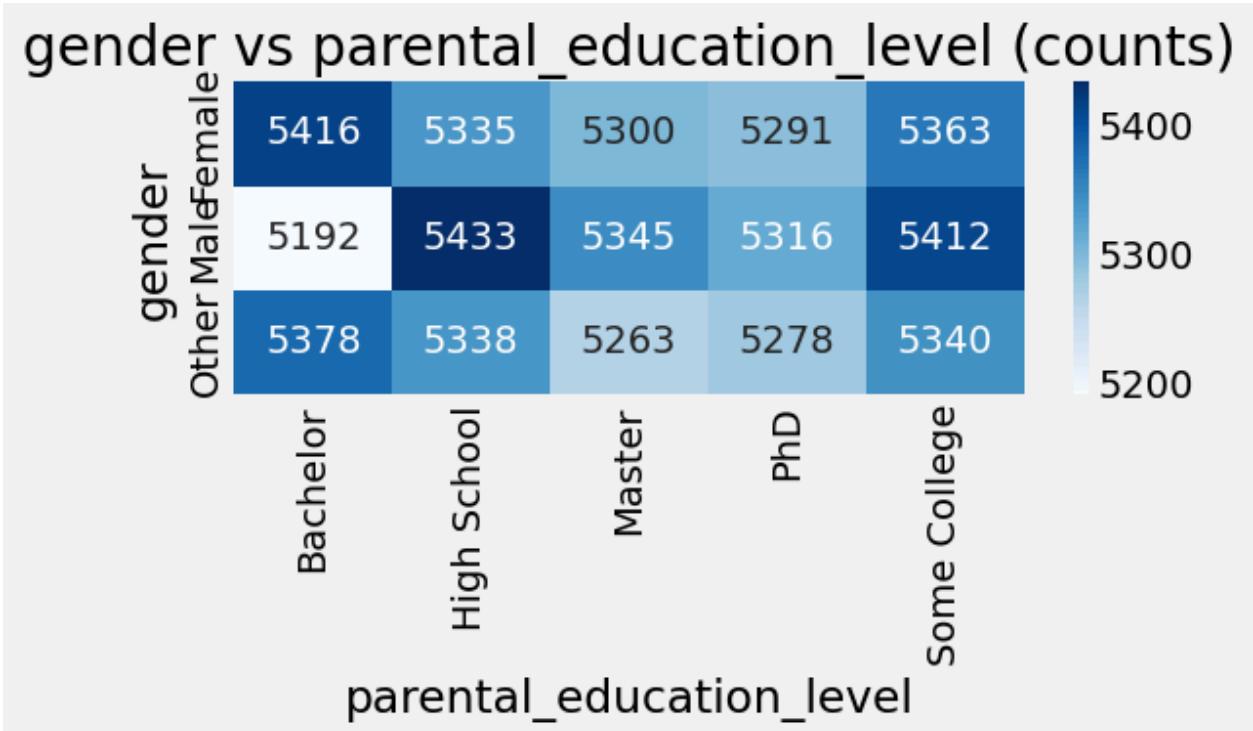
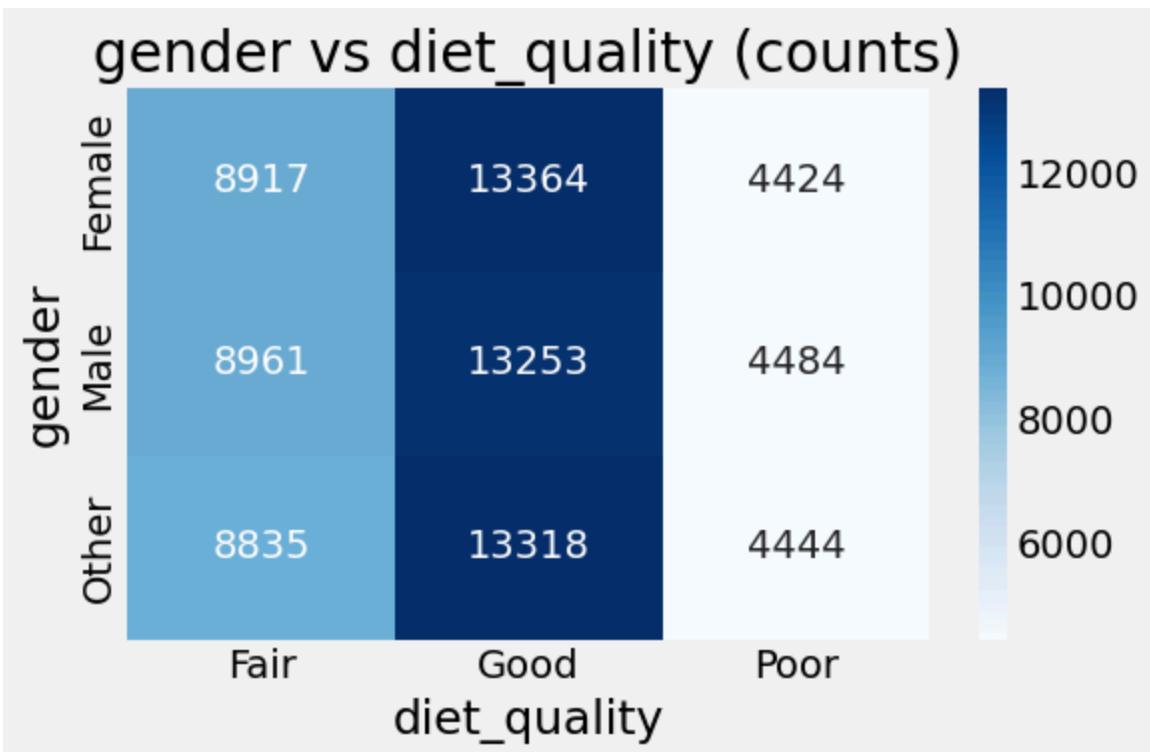








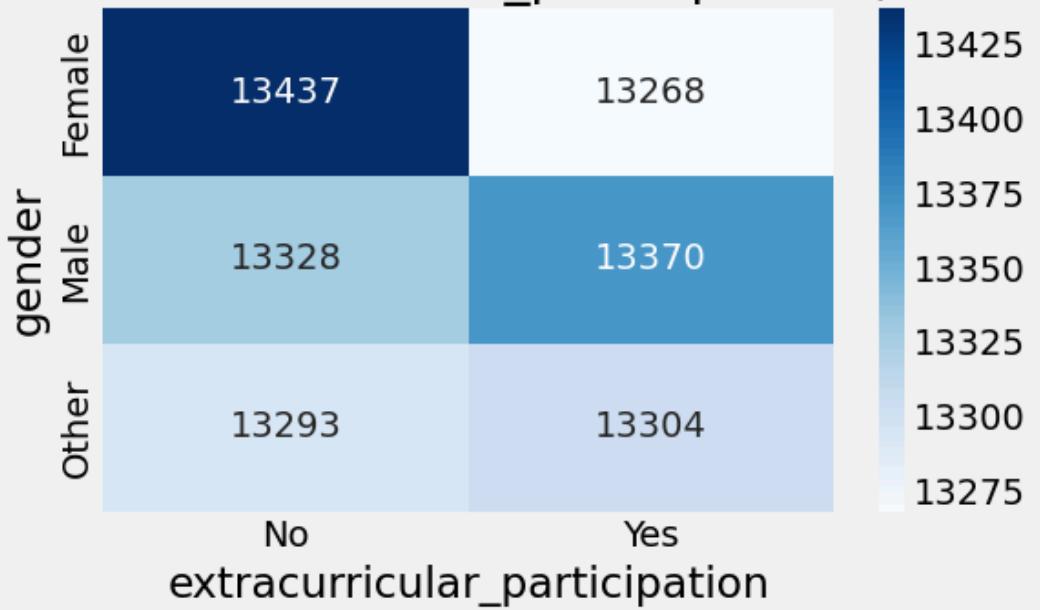




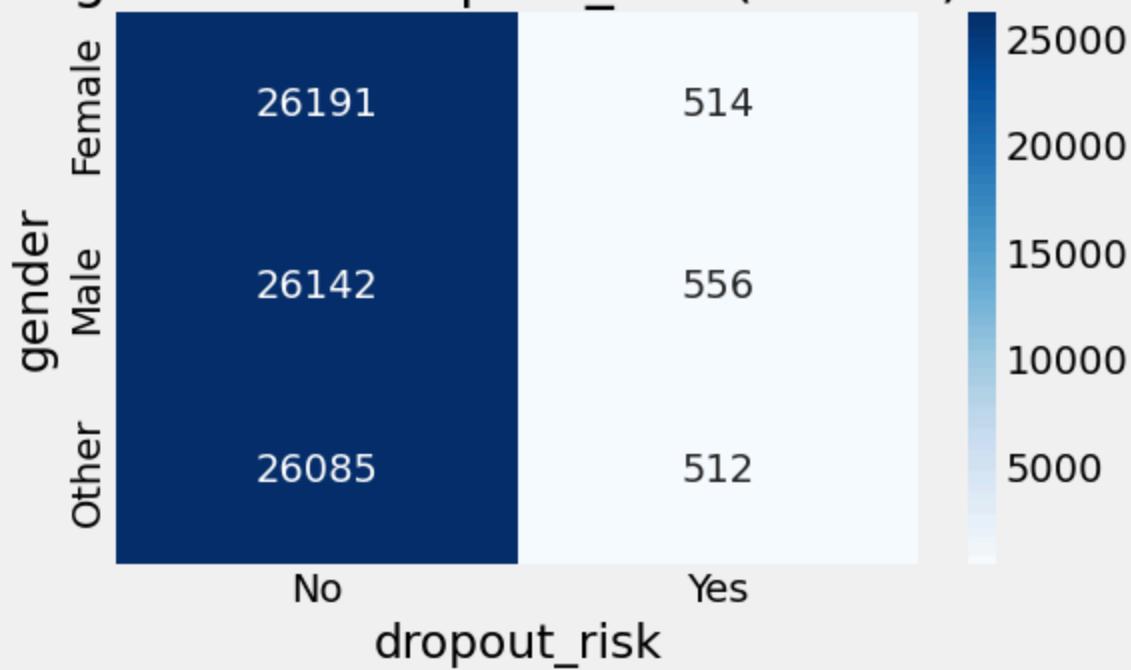
gender vs internet_quality (counts)



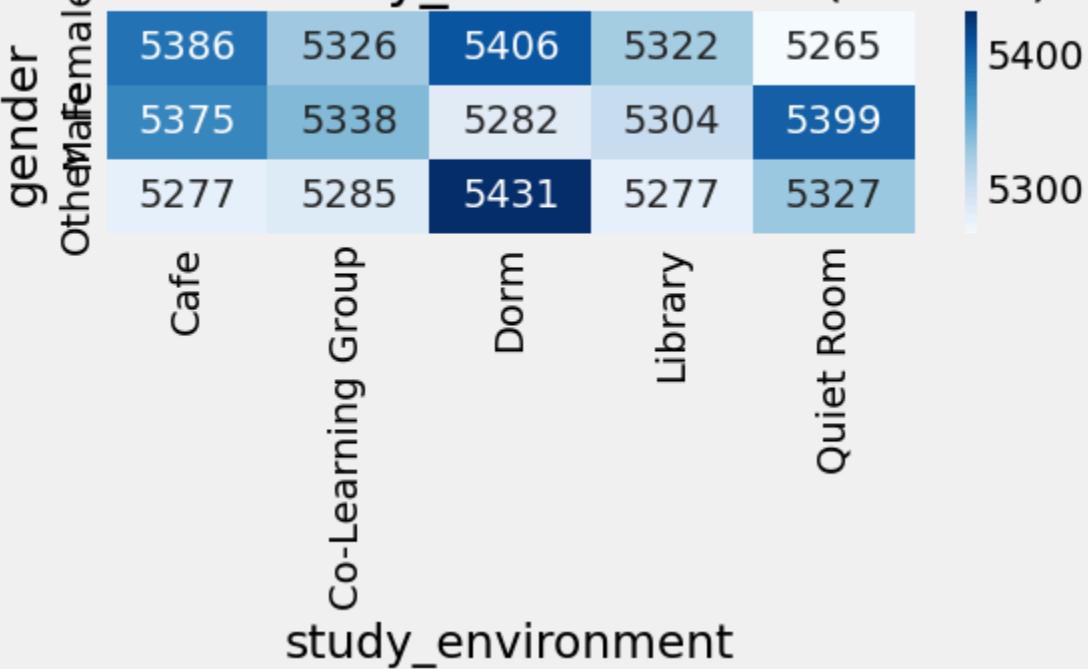
gender vs extracurricular_participation (counts)



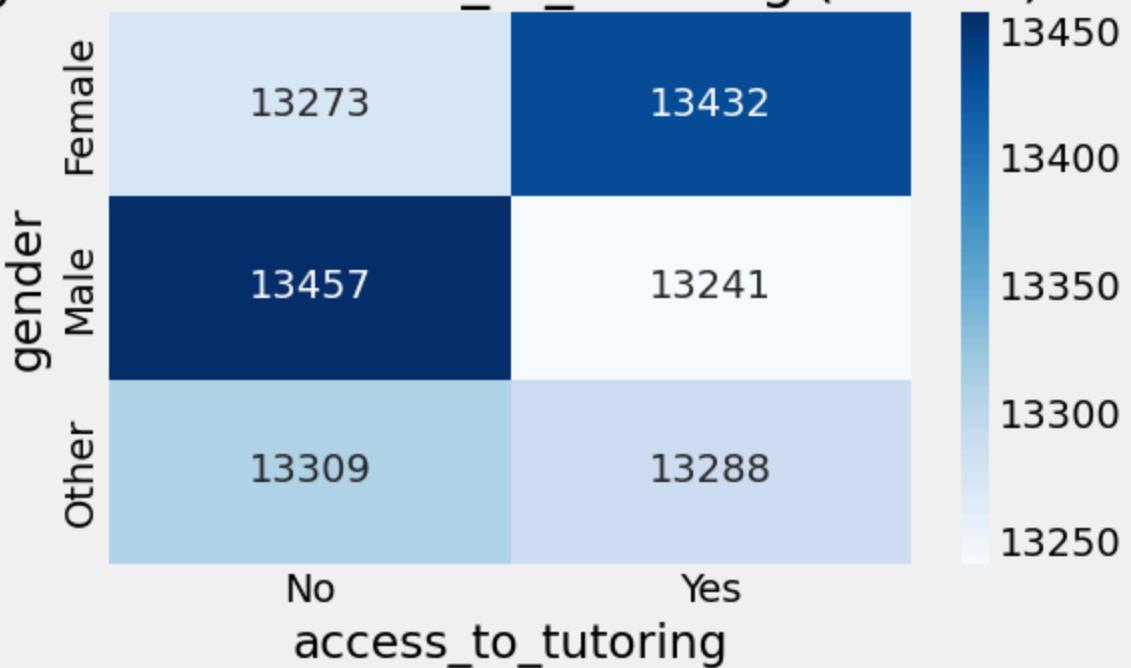
gender vs dropout_risk (counts)



gender vs study_environment (counts)



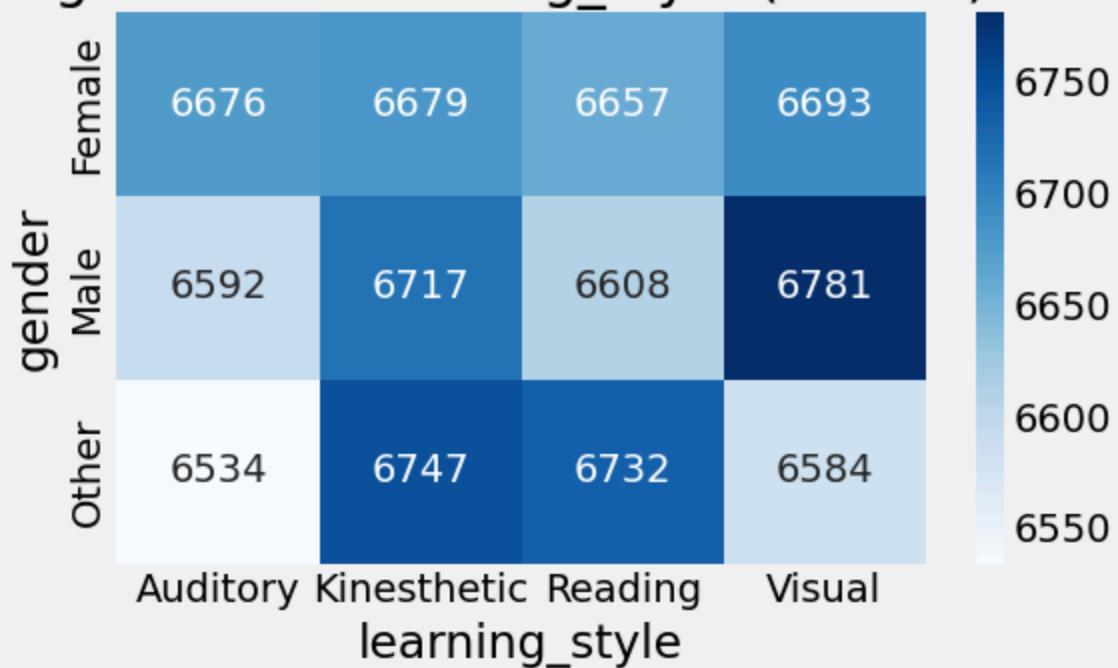
gender vs access_to_tutoring (counts)



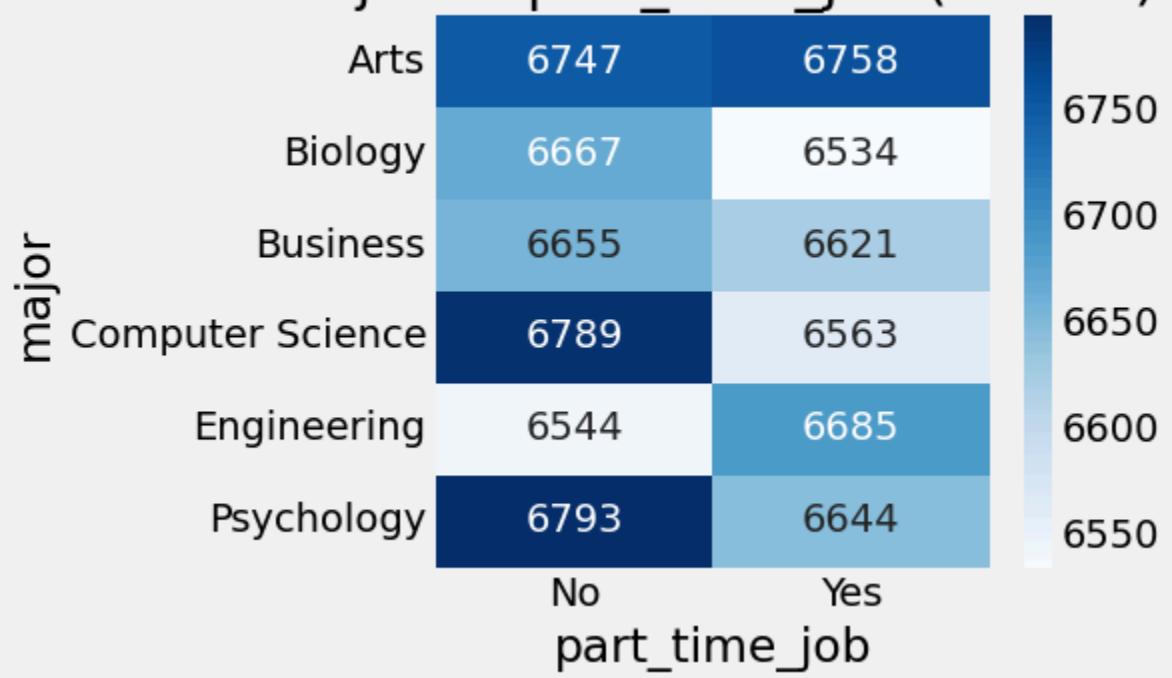
gender vs family_income_range (counts)

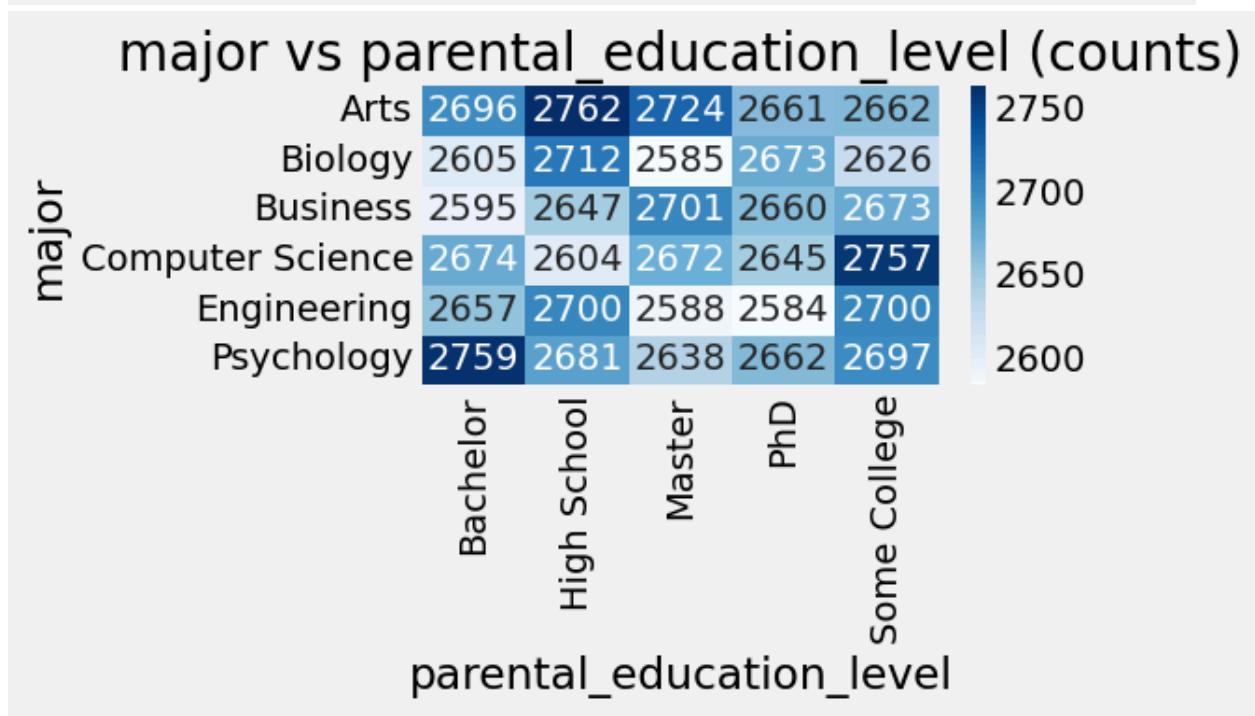
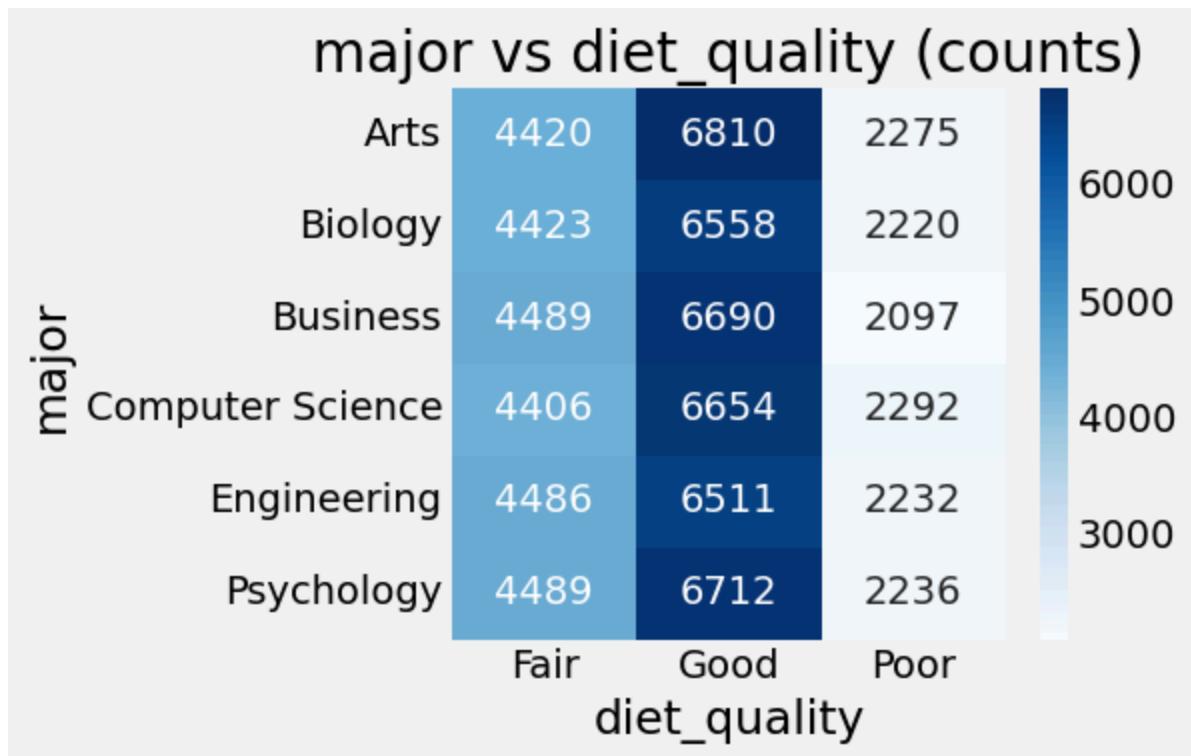


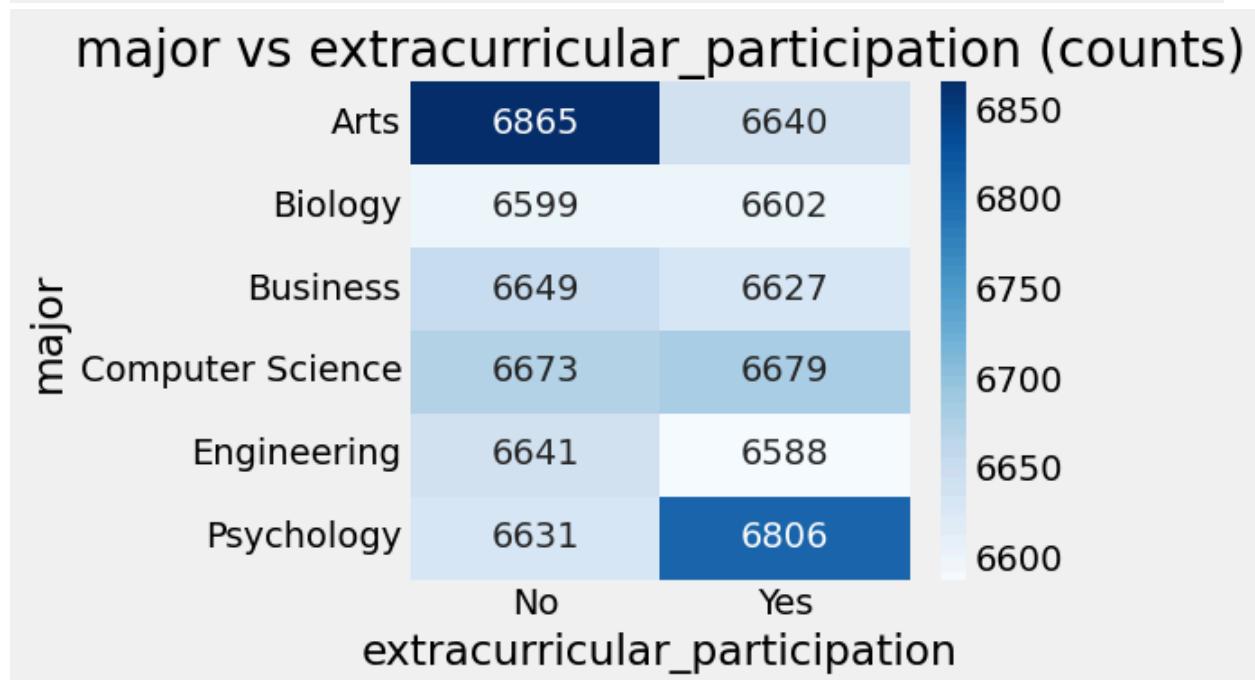
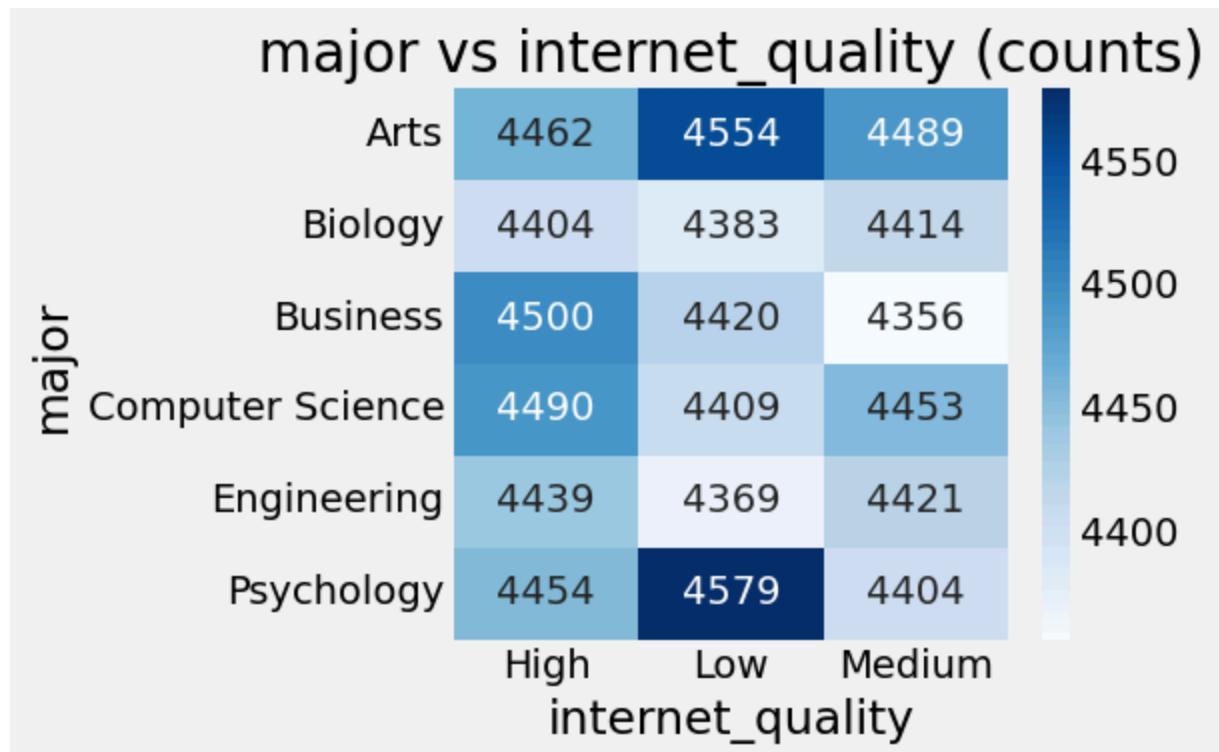
gender vs learning_style (counts)

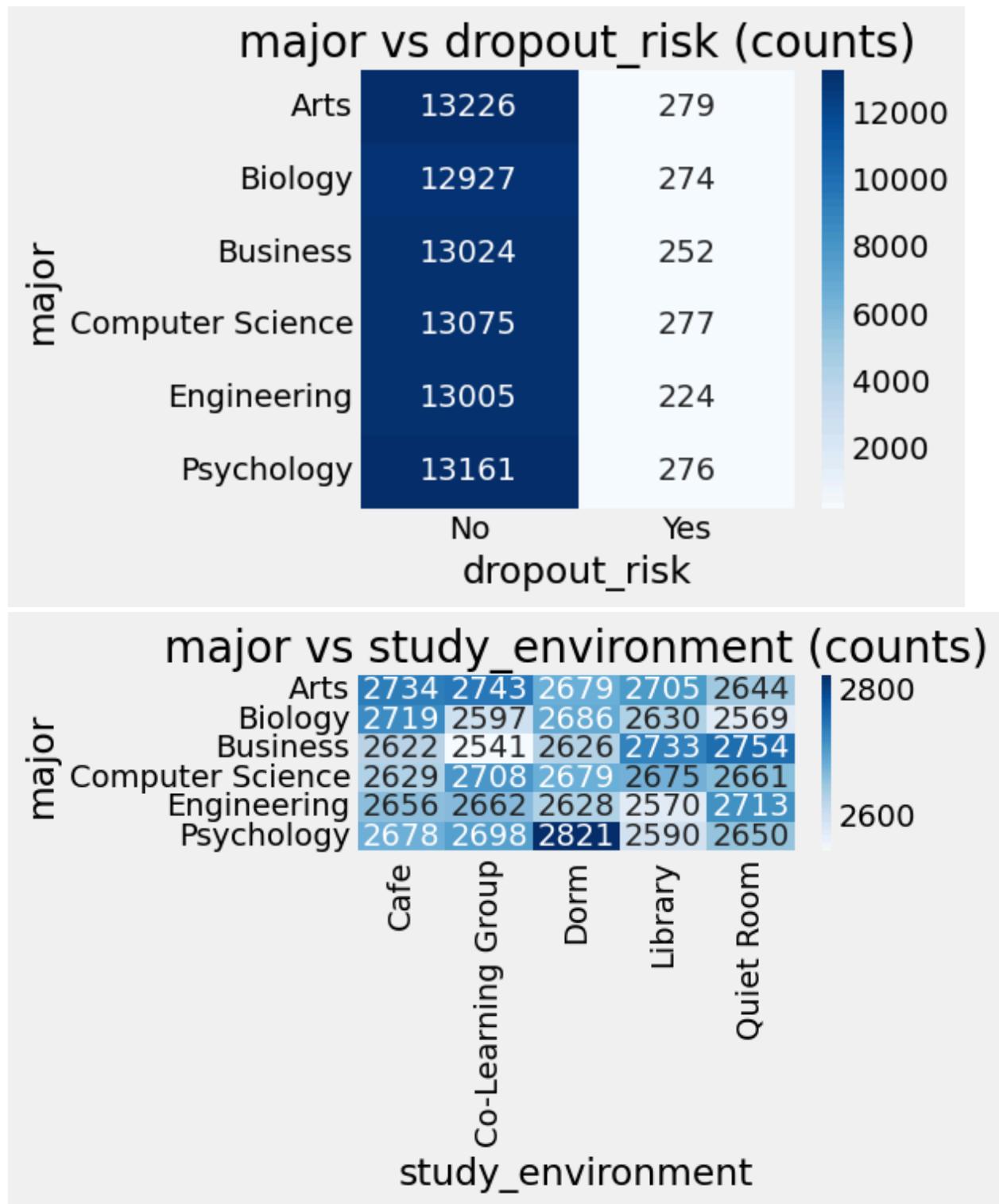


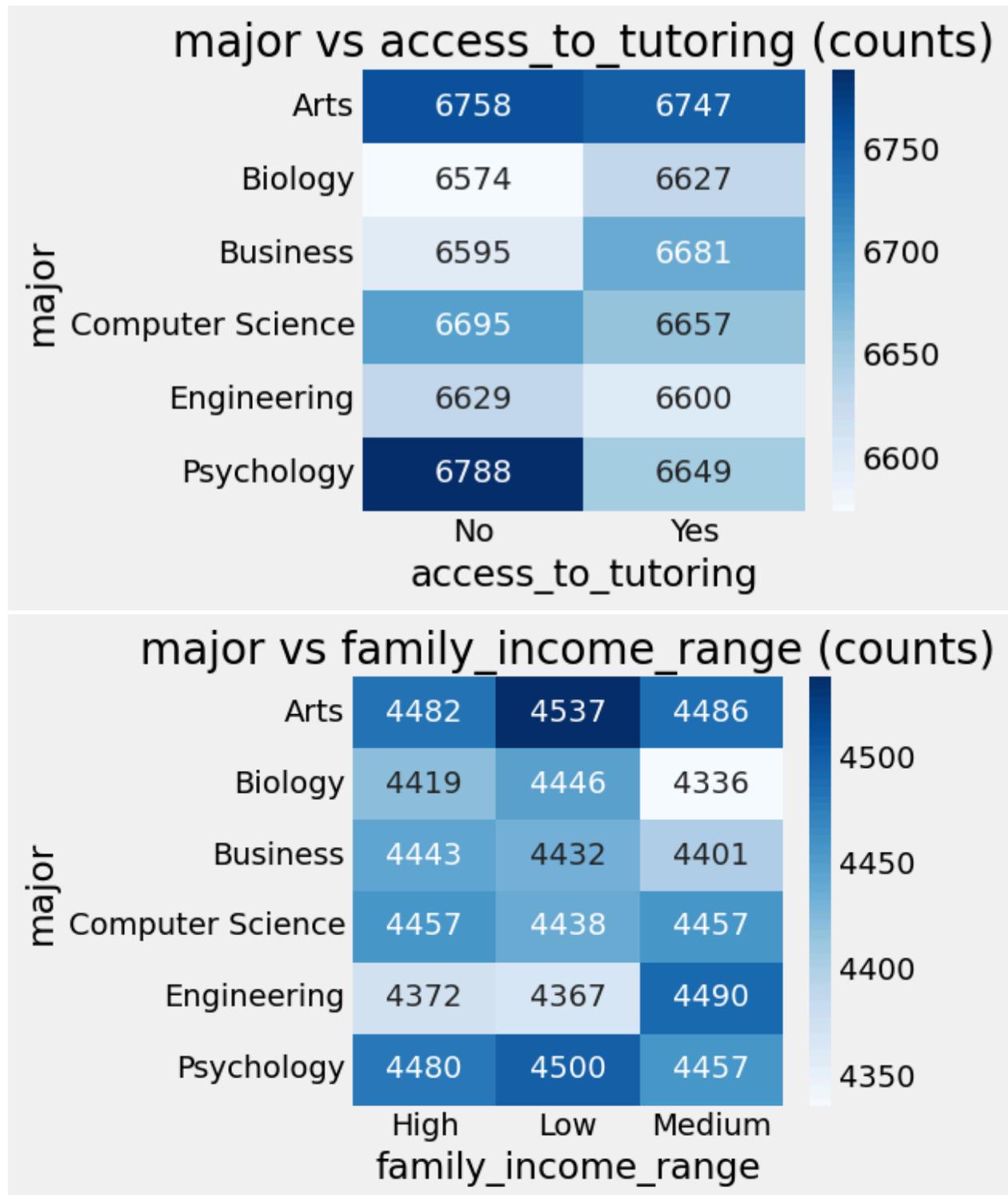
major vs part_time_job (counts)

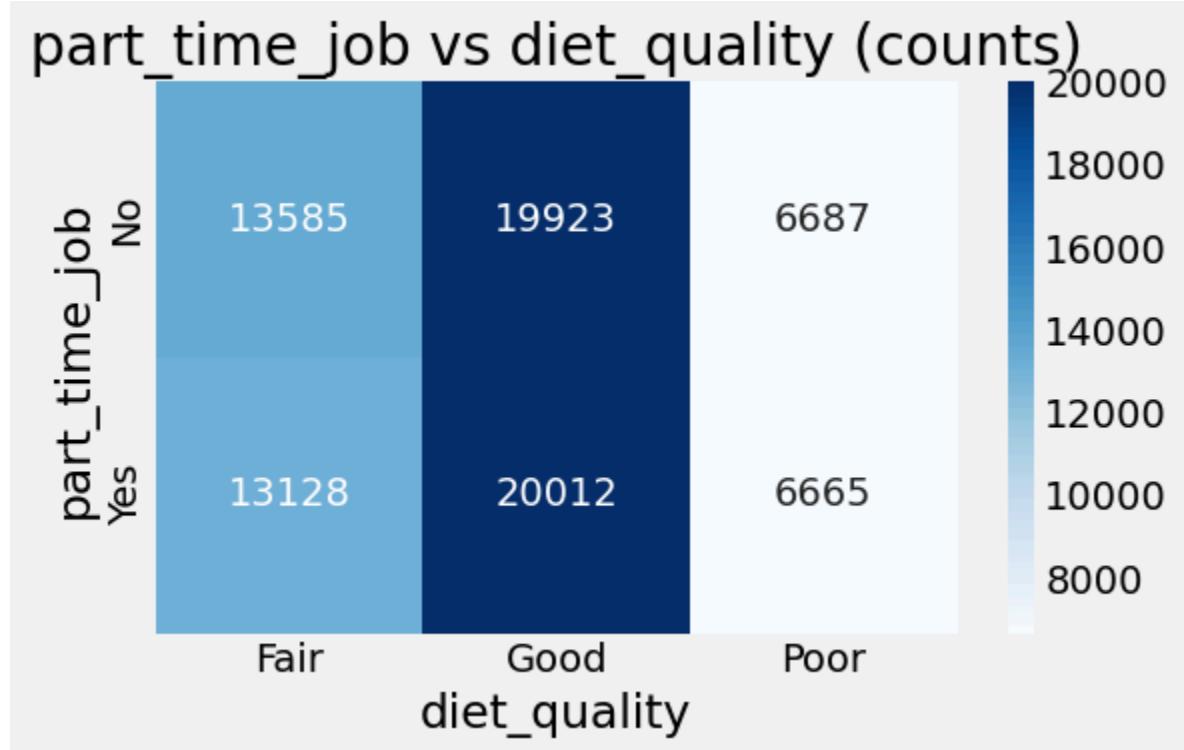
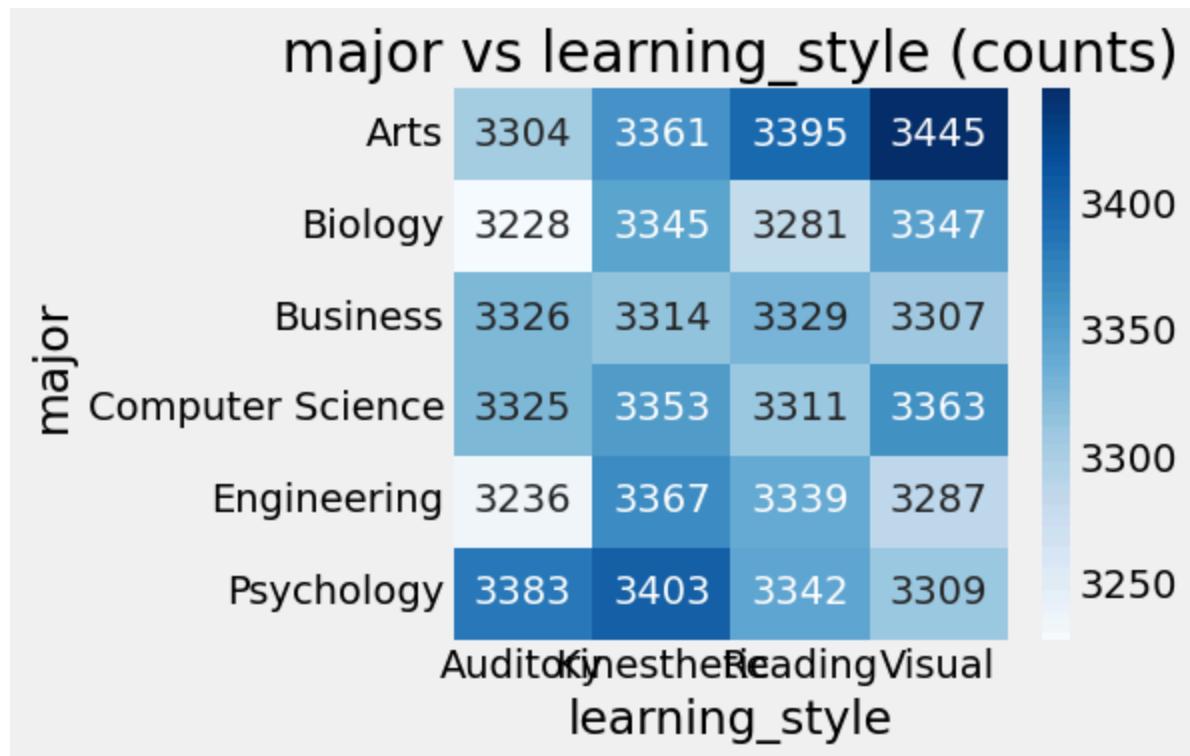


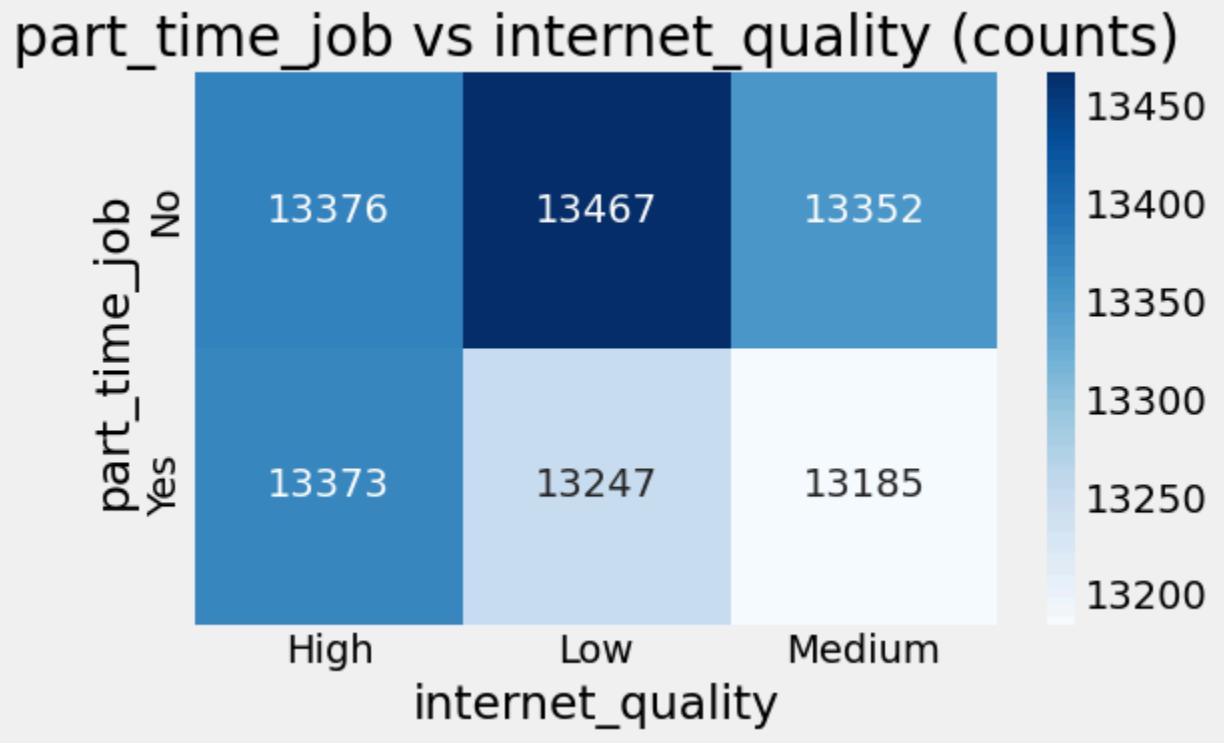
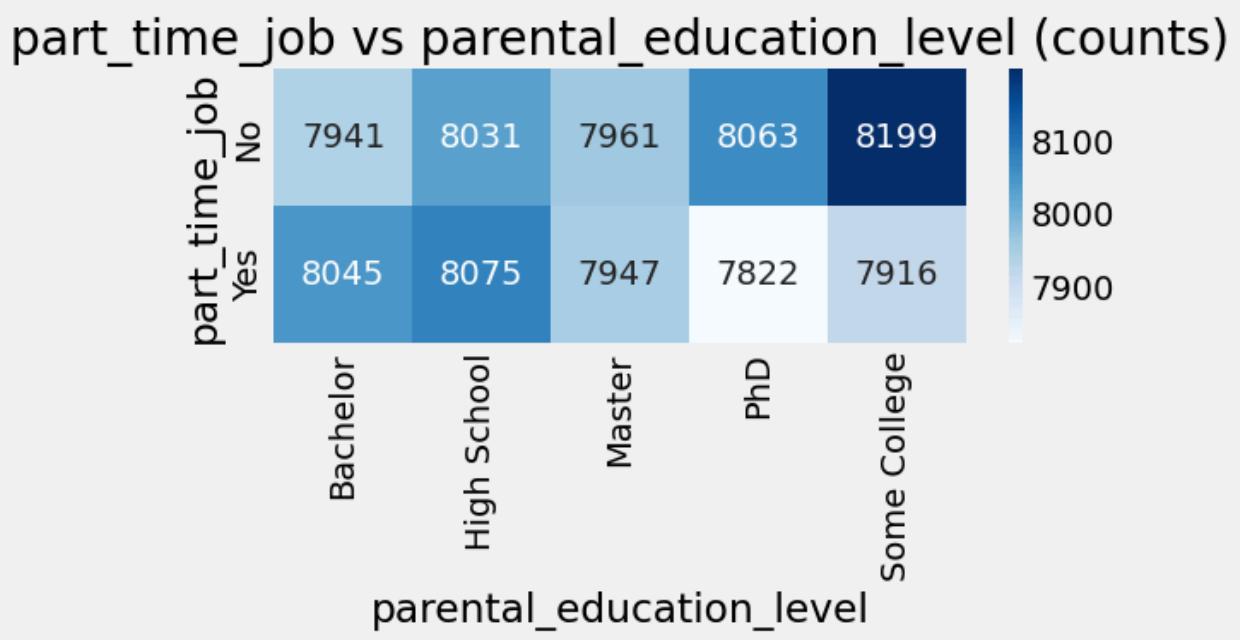


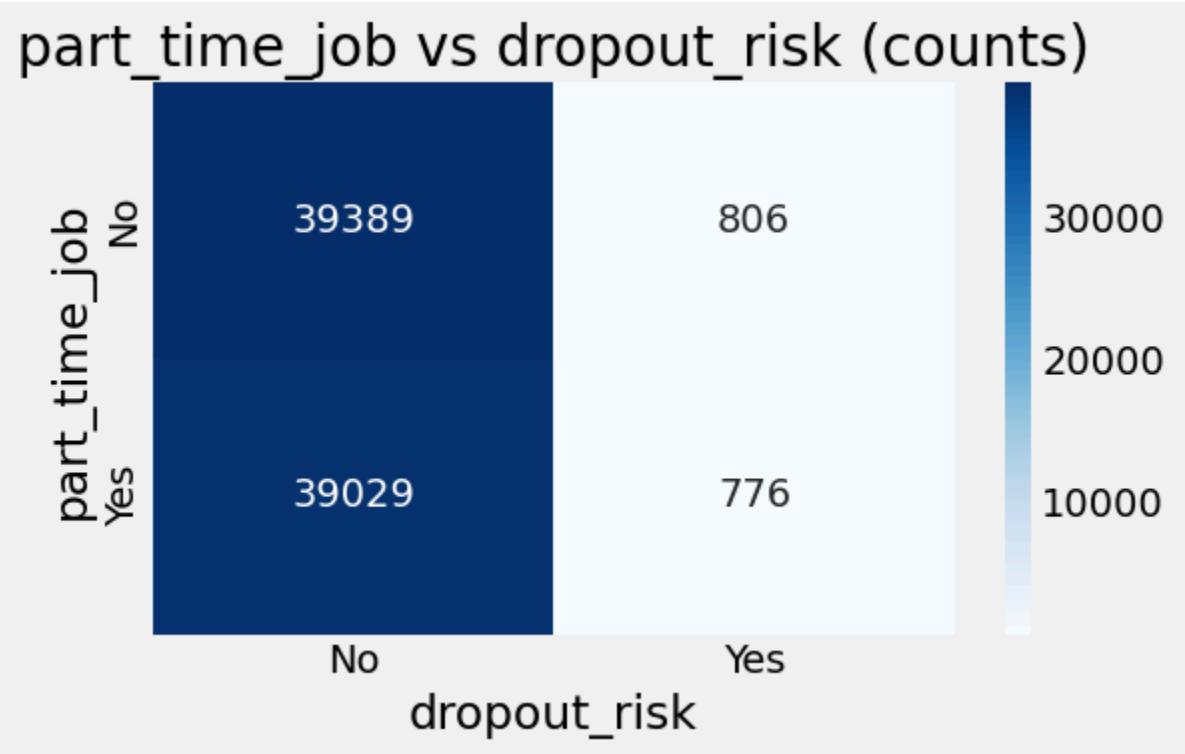
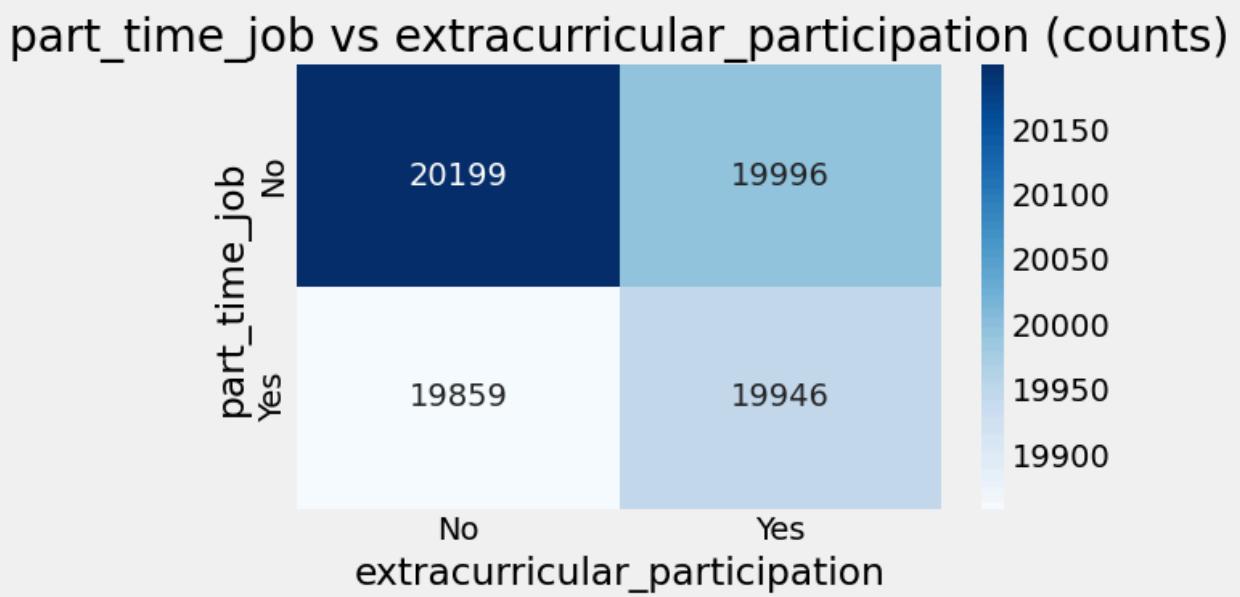


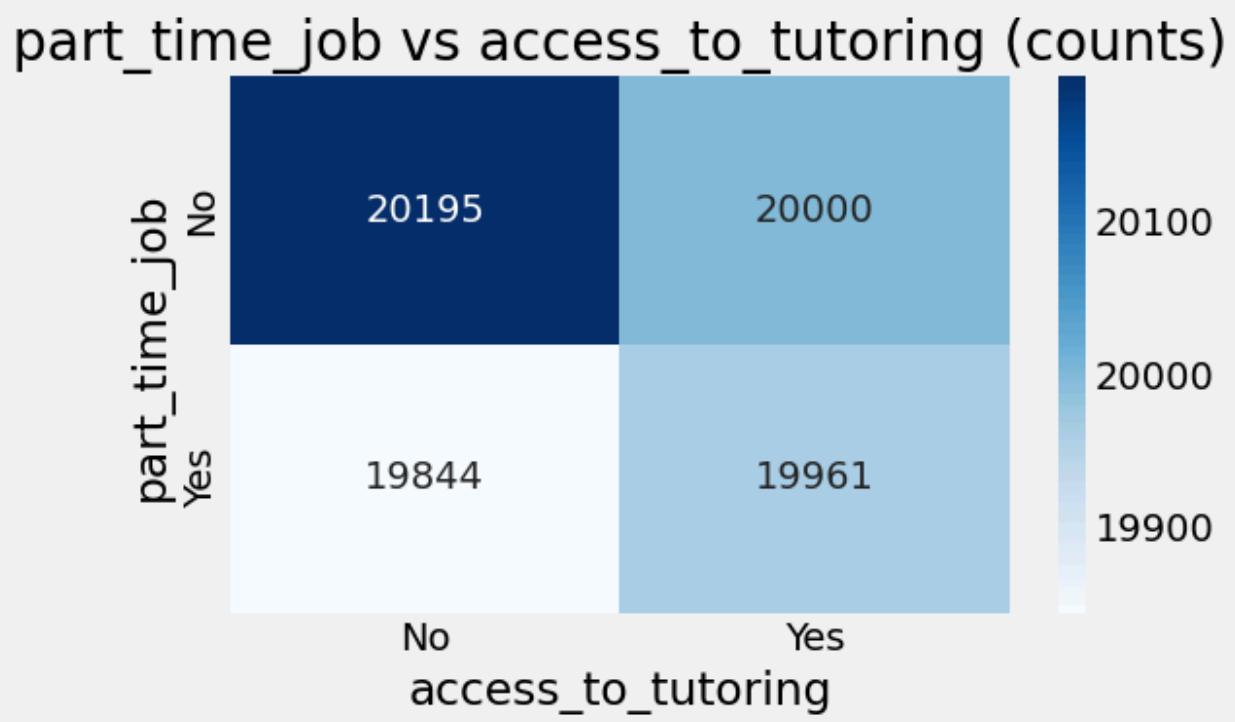
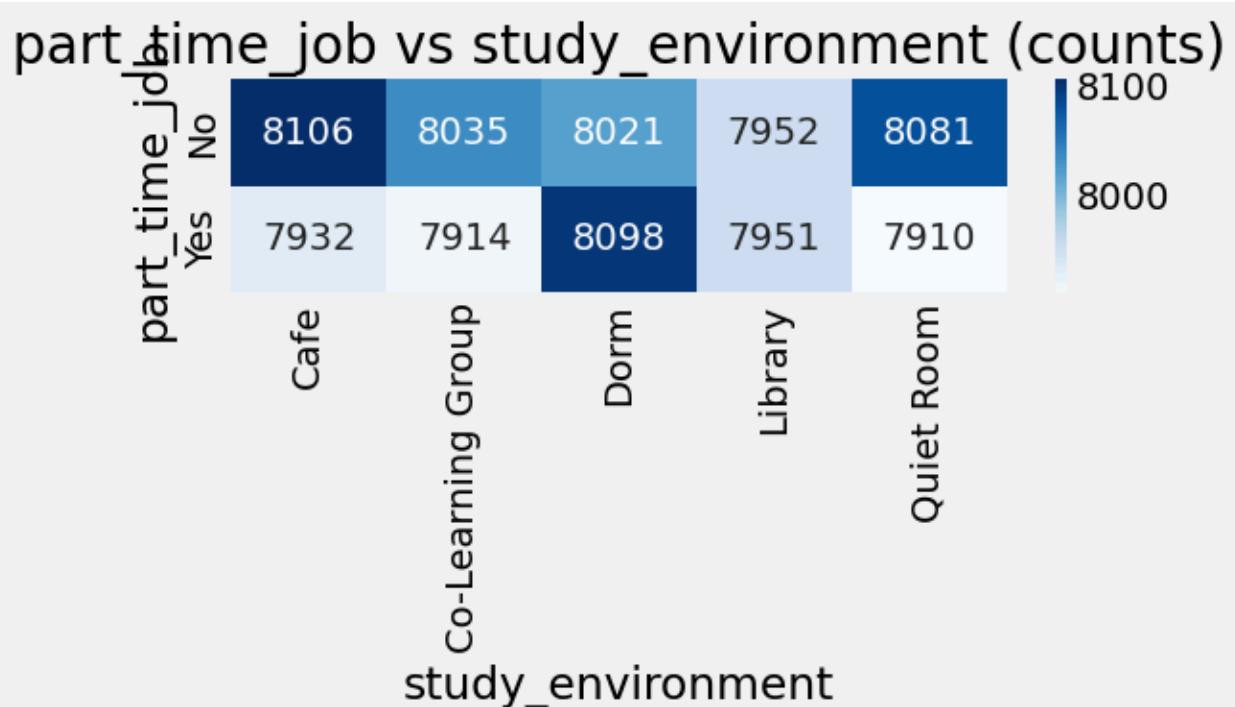


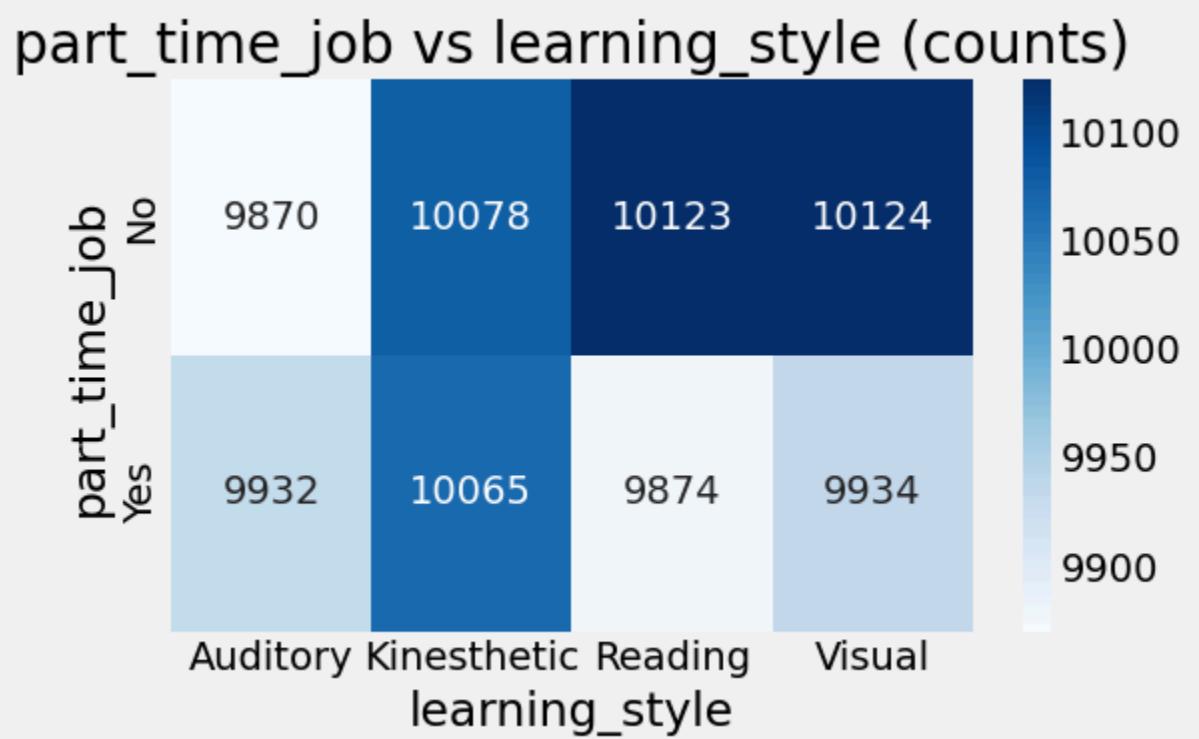
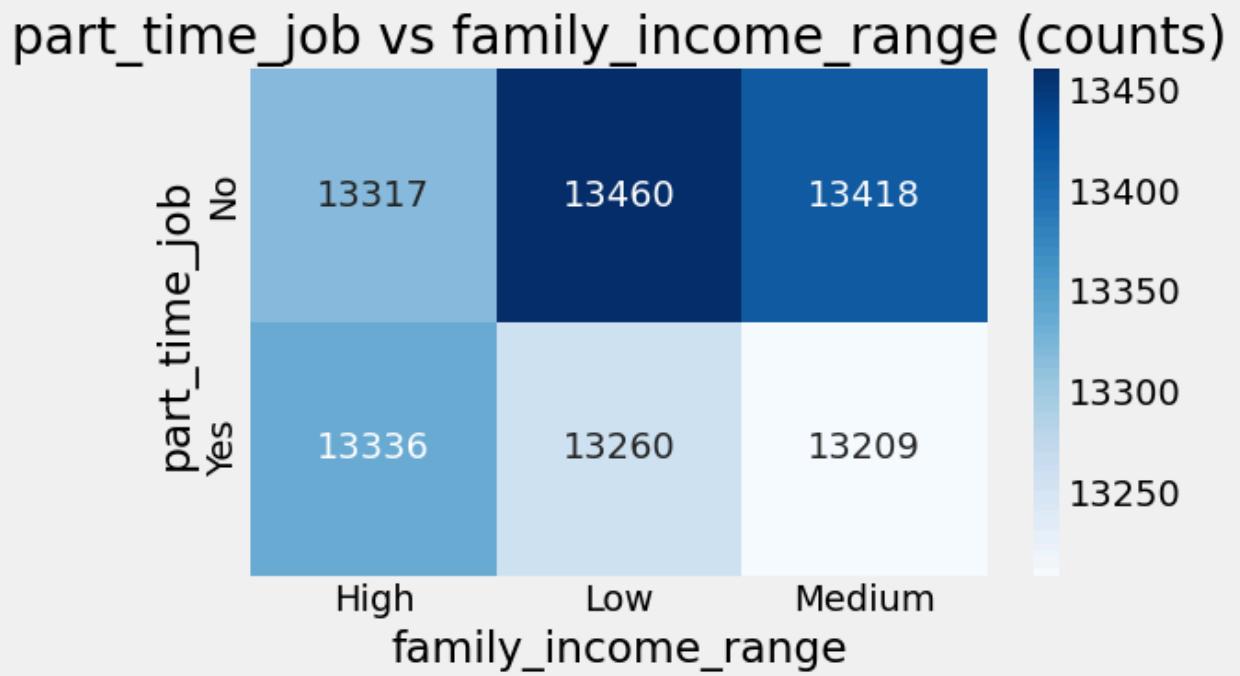


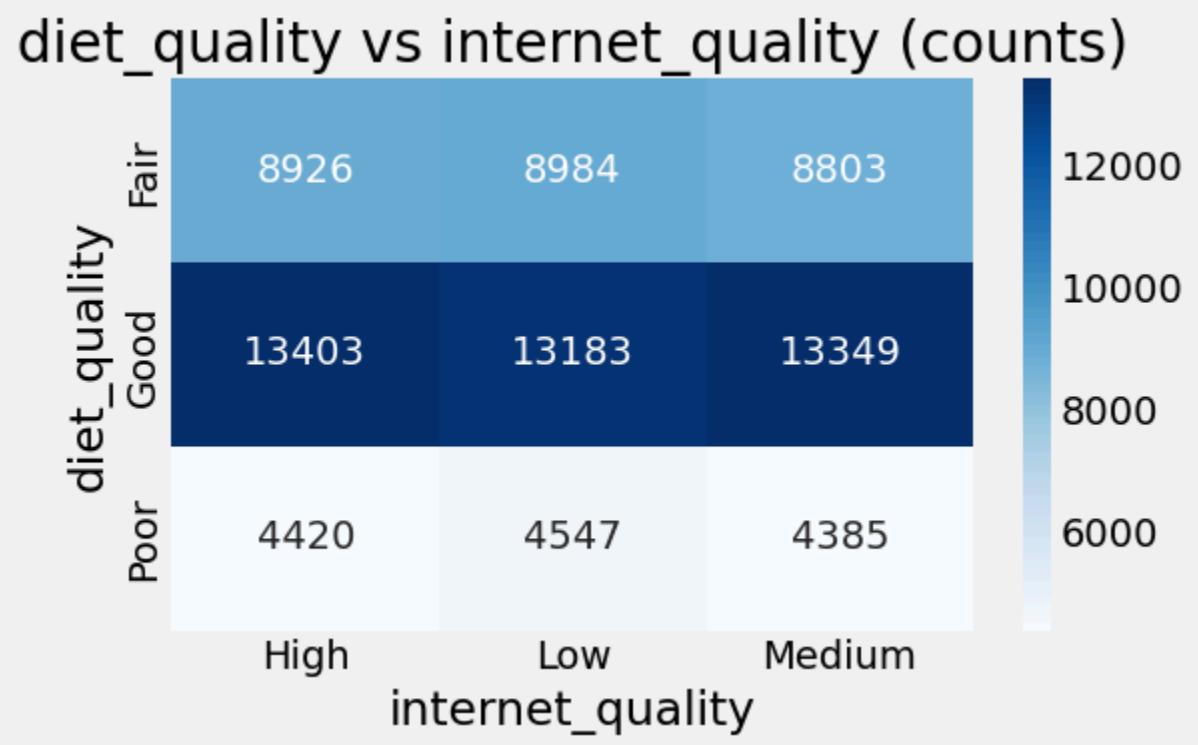
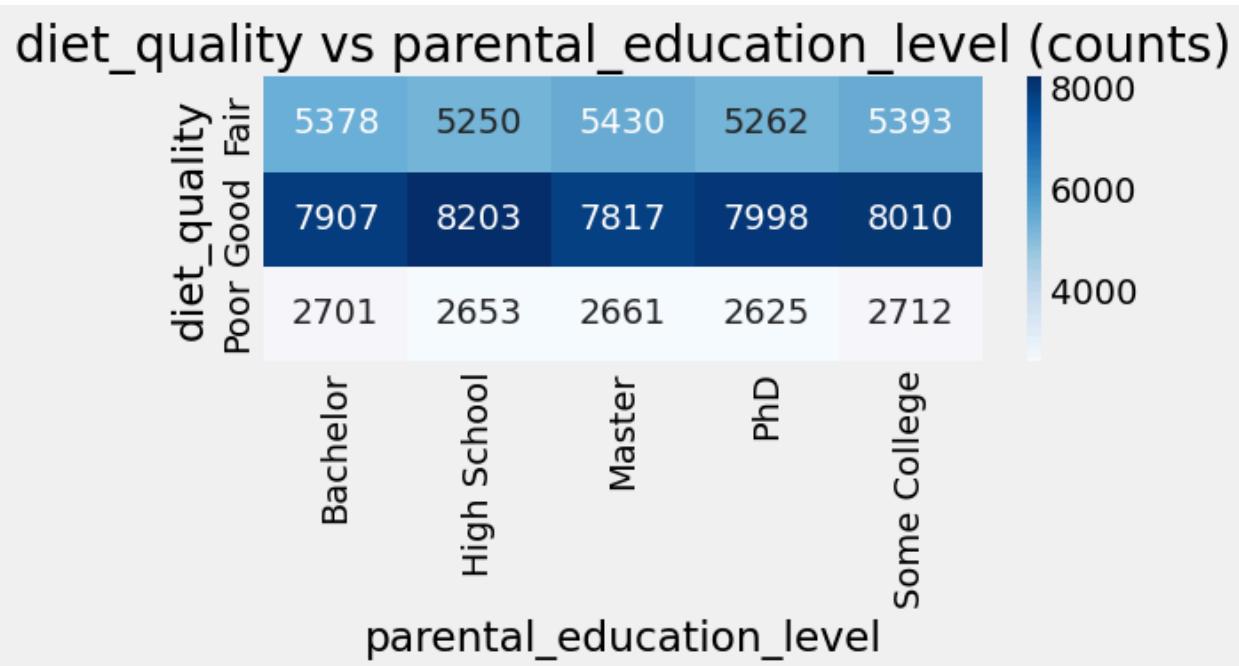


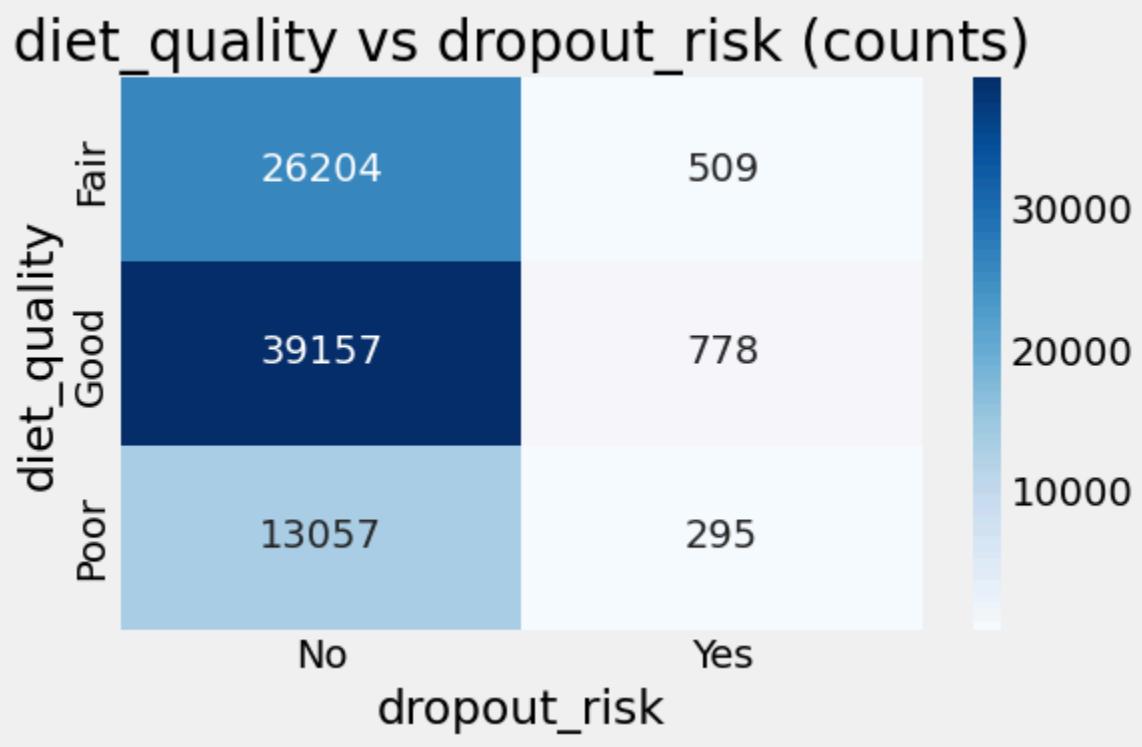
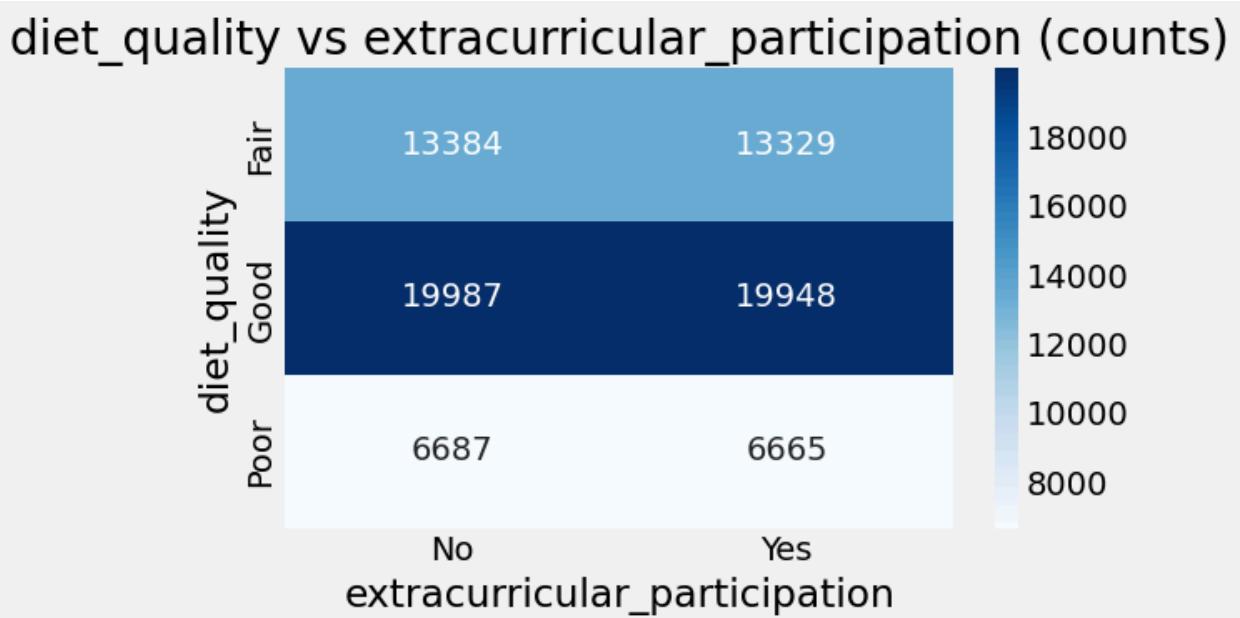


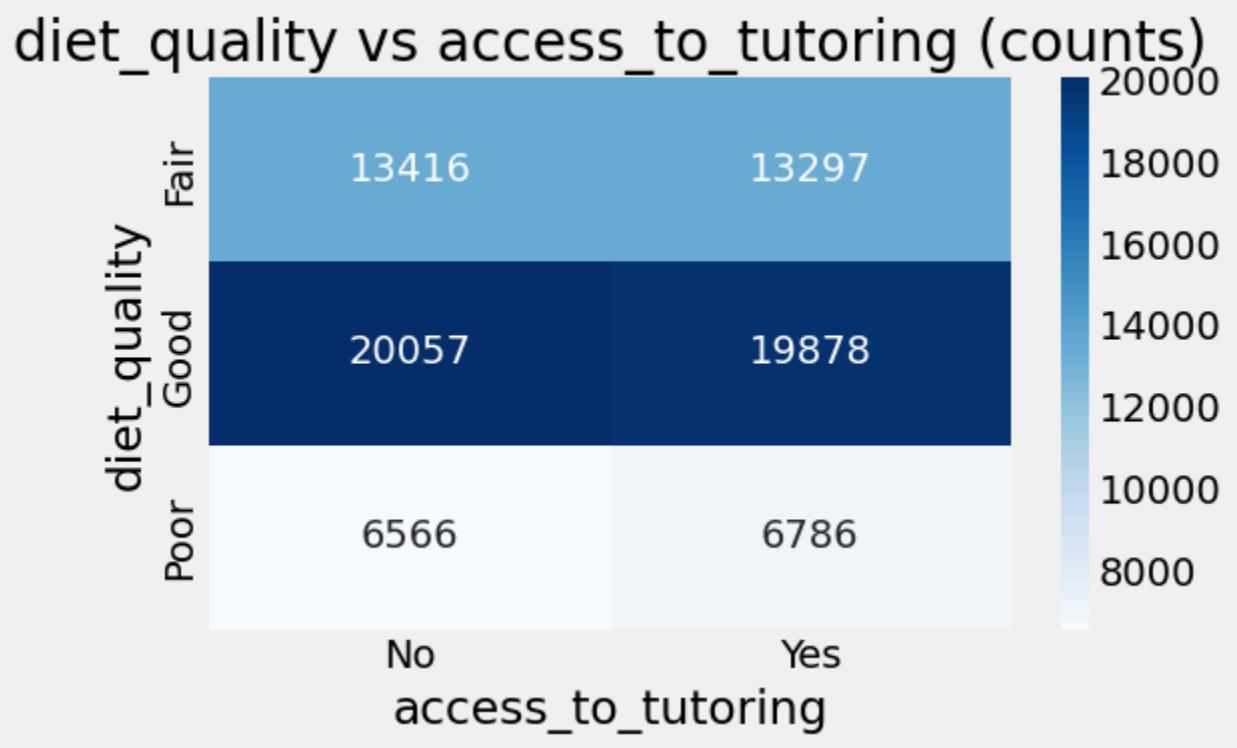
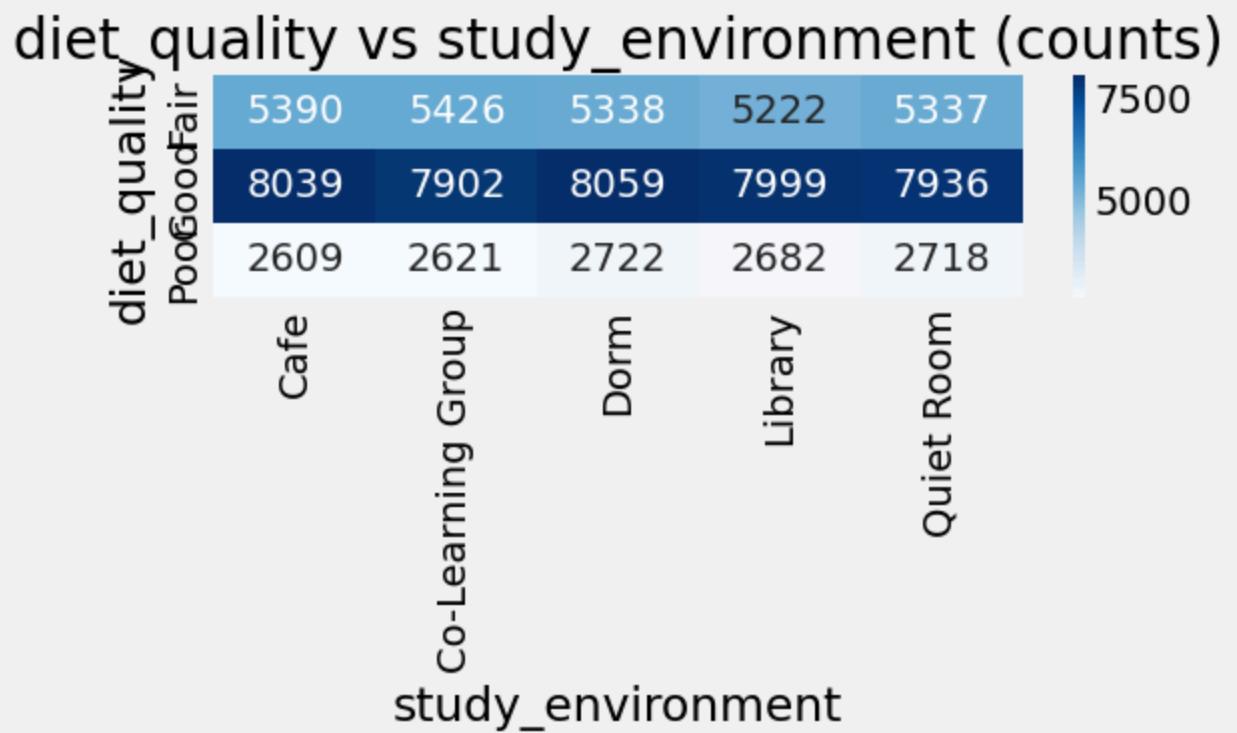


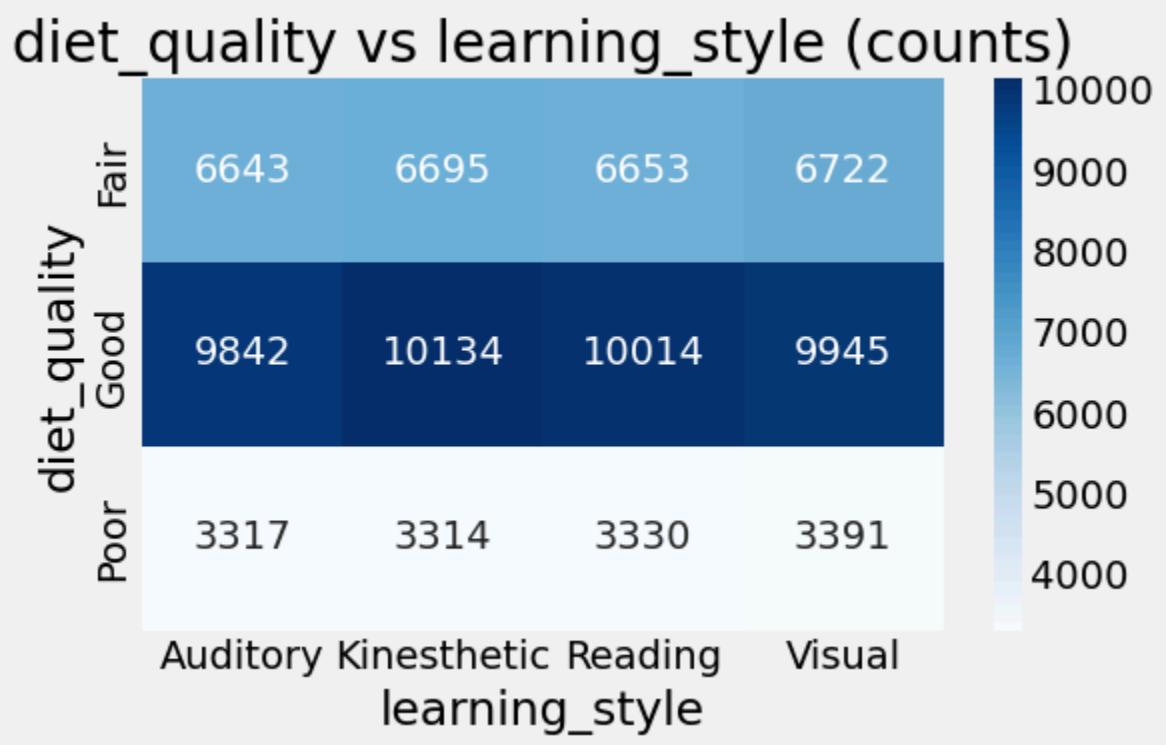
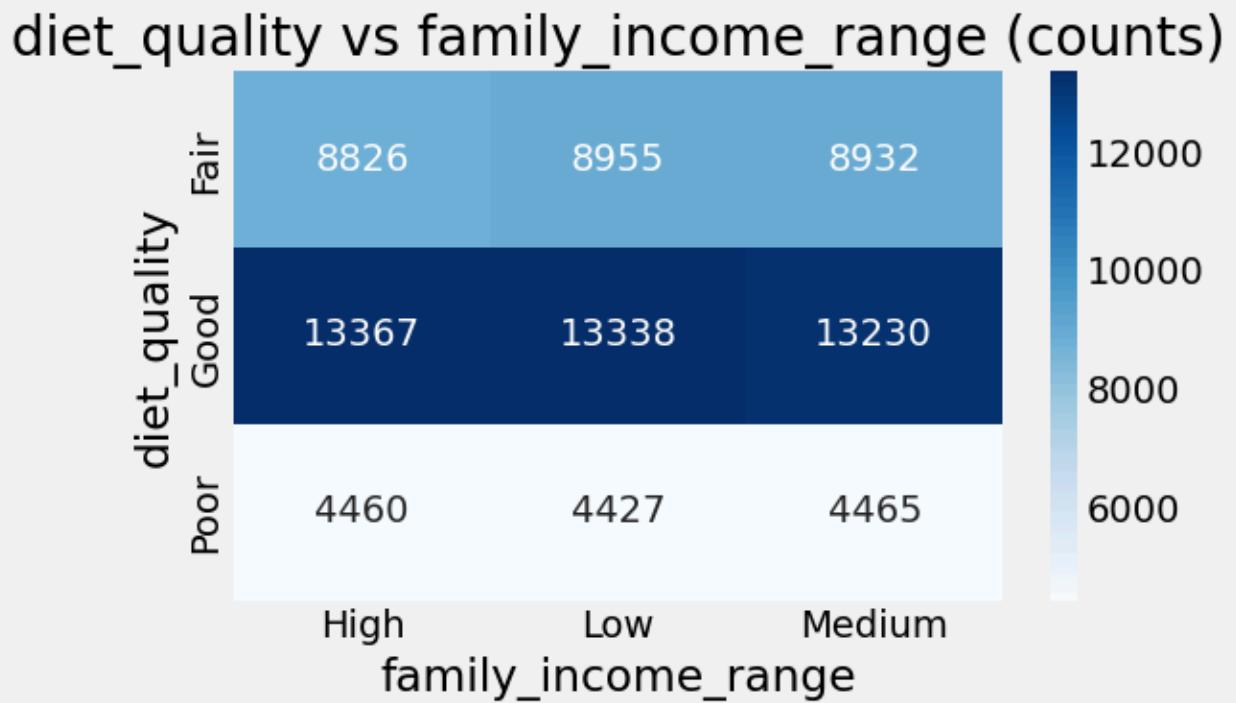




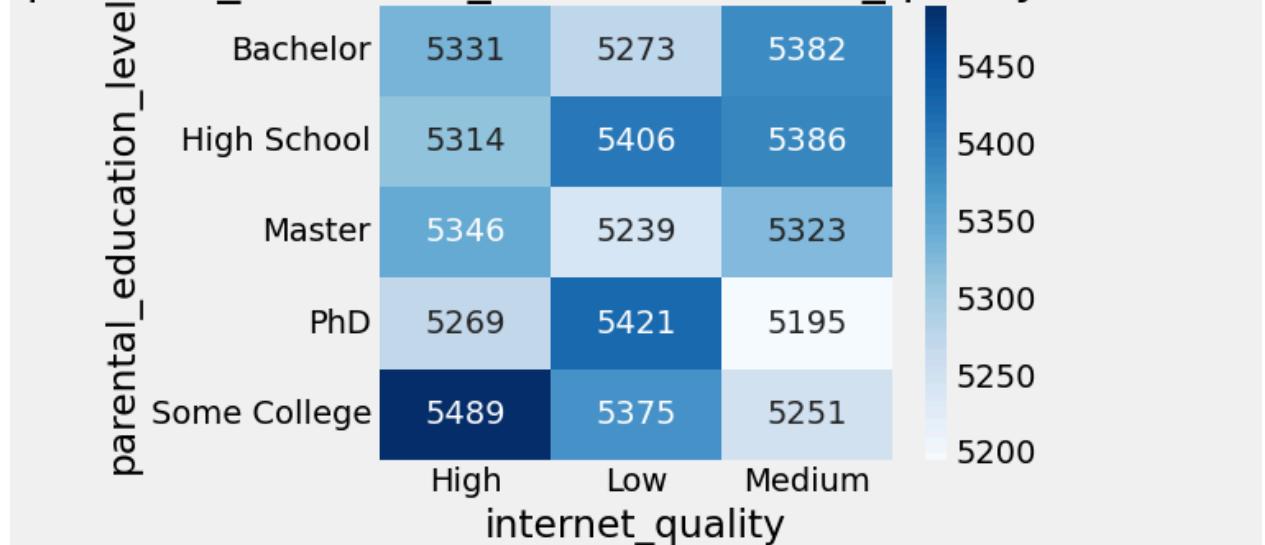




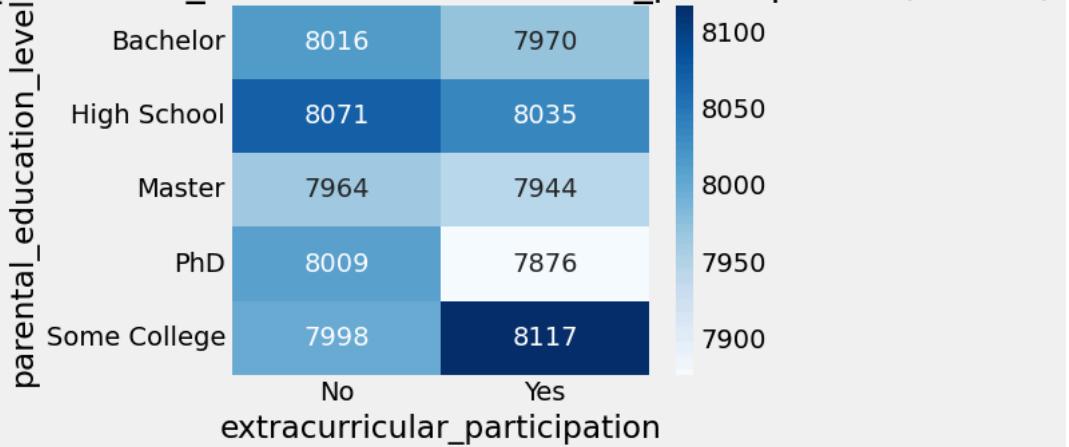




parental_education_level vs internet_quality (counts)

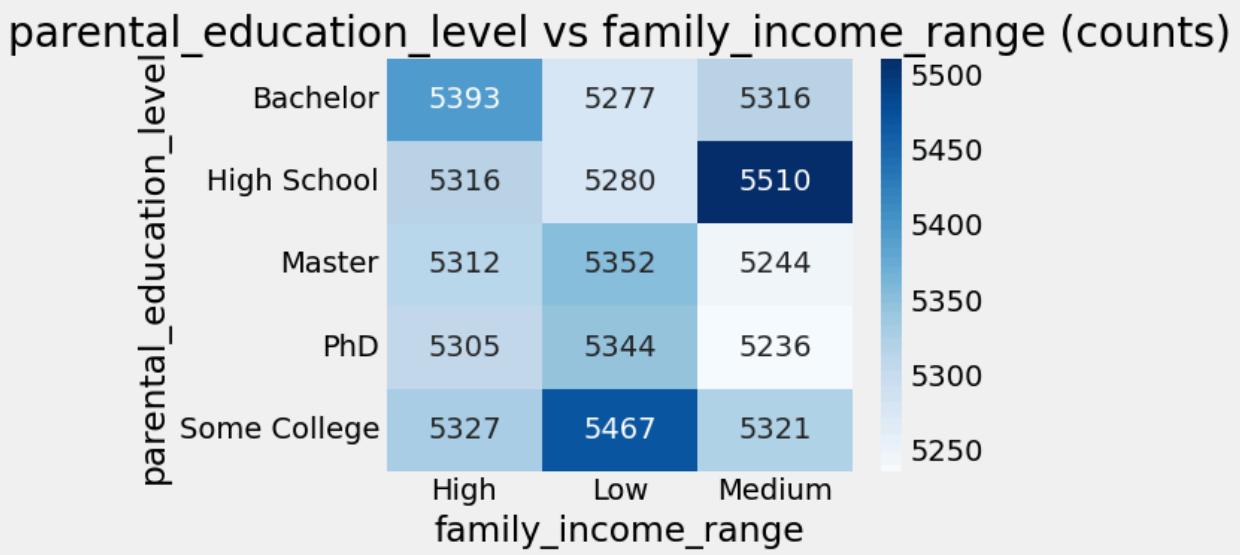
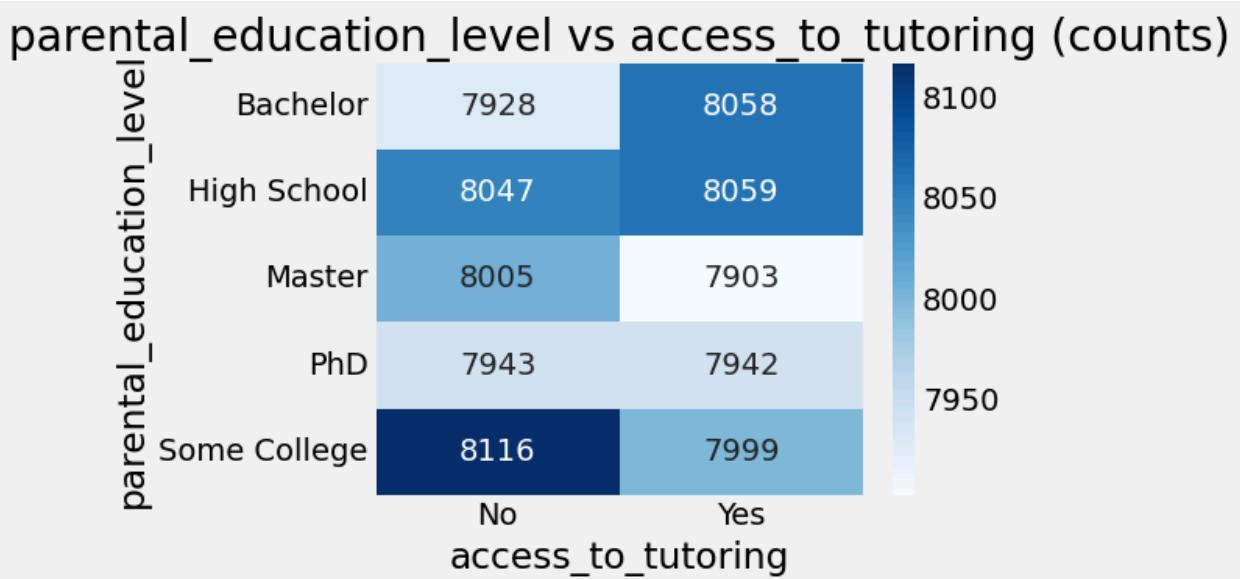
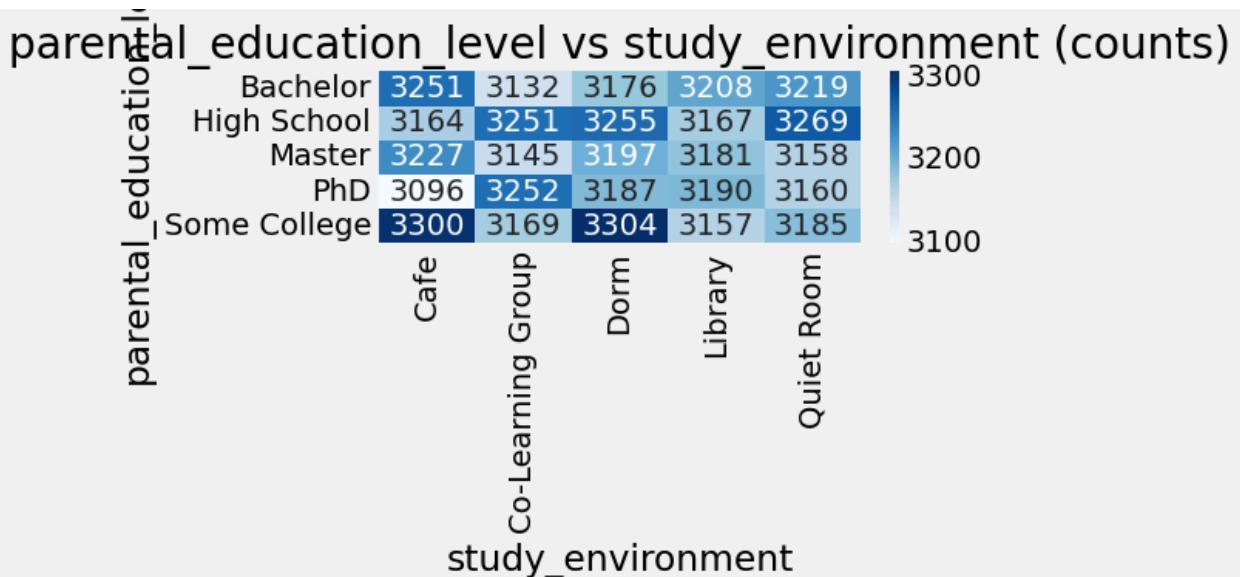


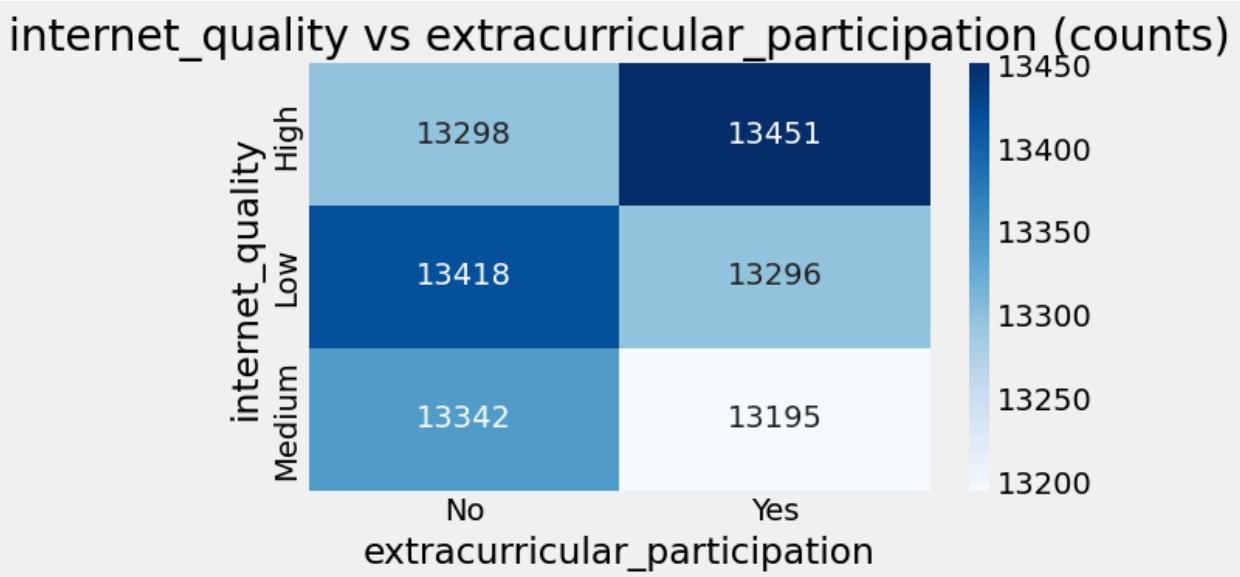
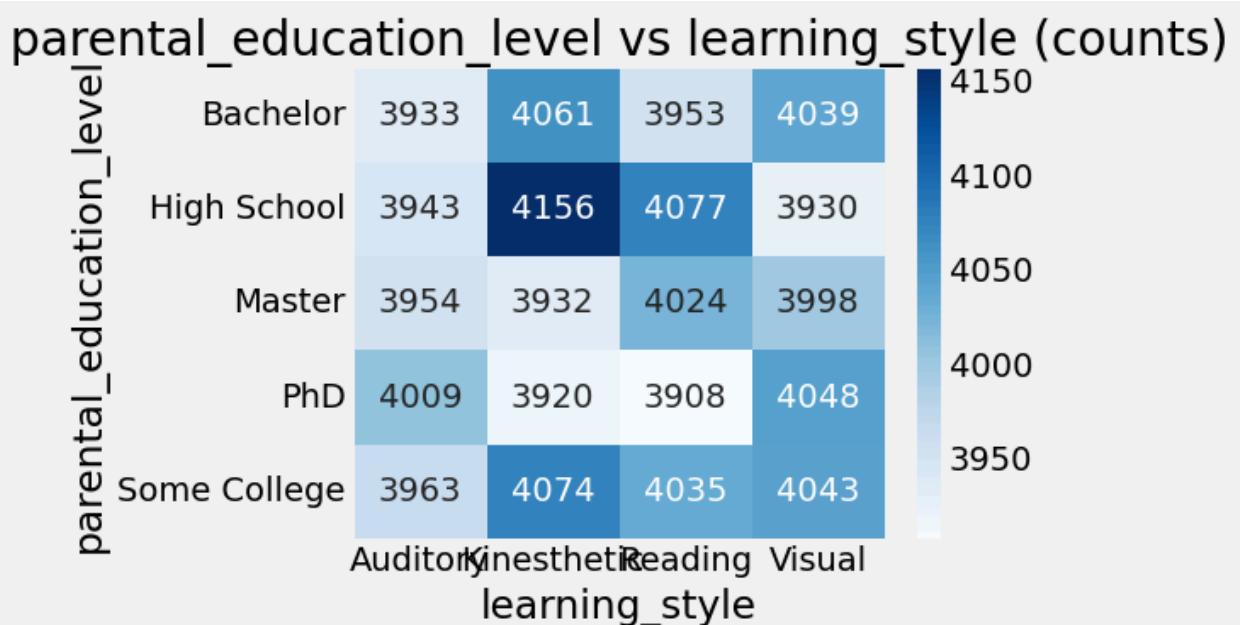
parental_education_level vs extracurricular_participation (counts)



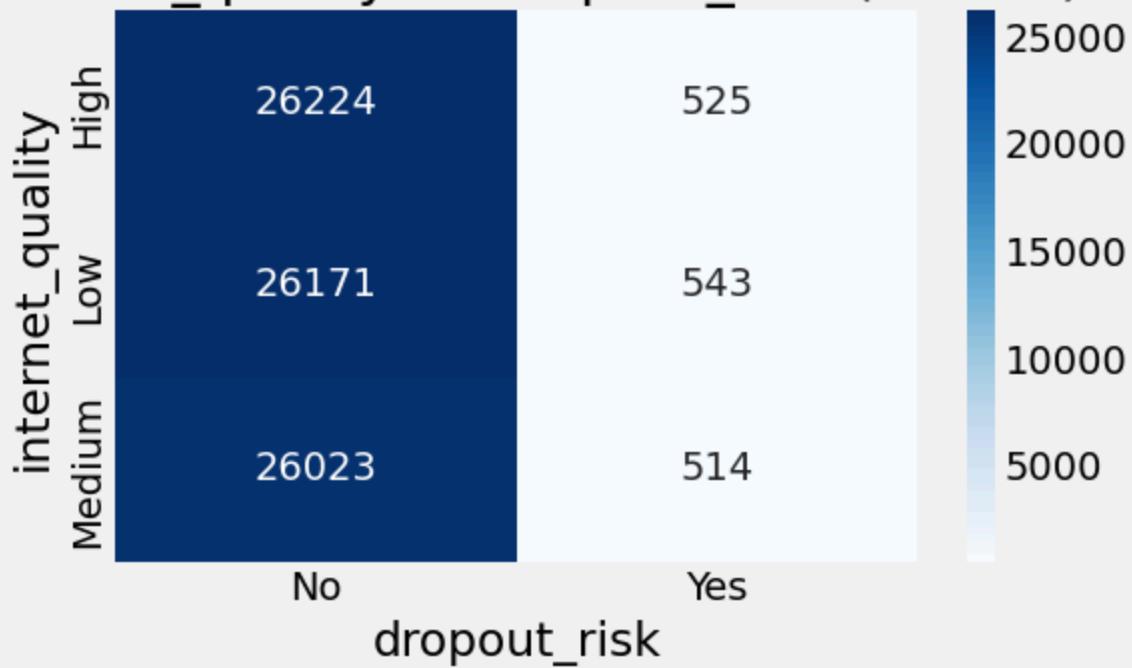
parental_education_level vs dropout_risk (counts)



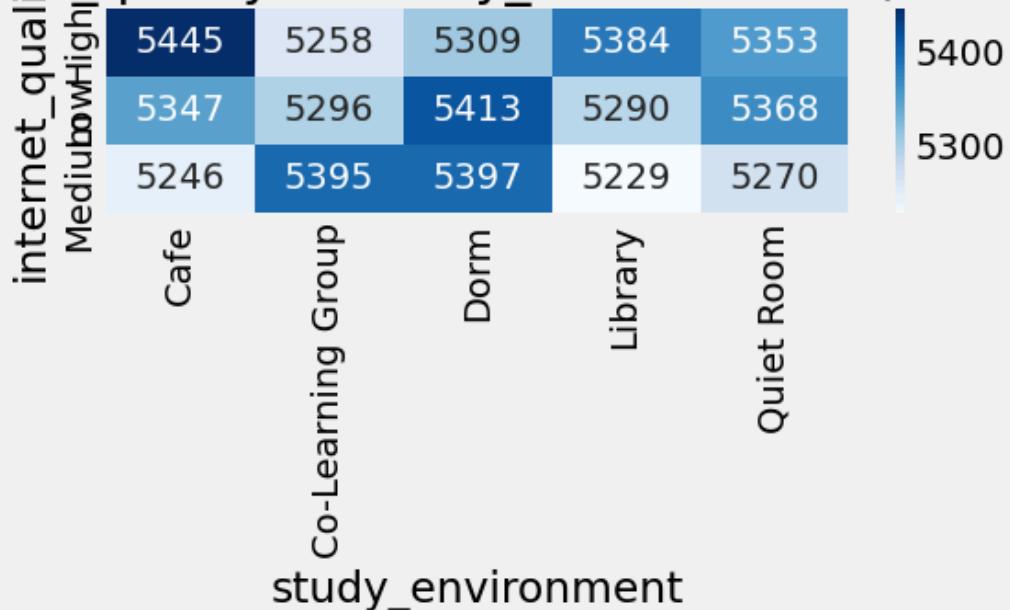


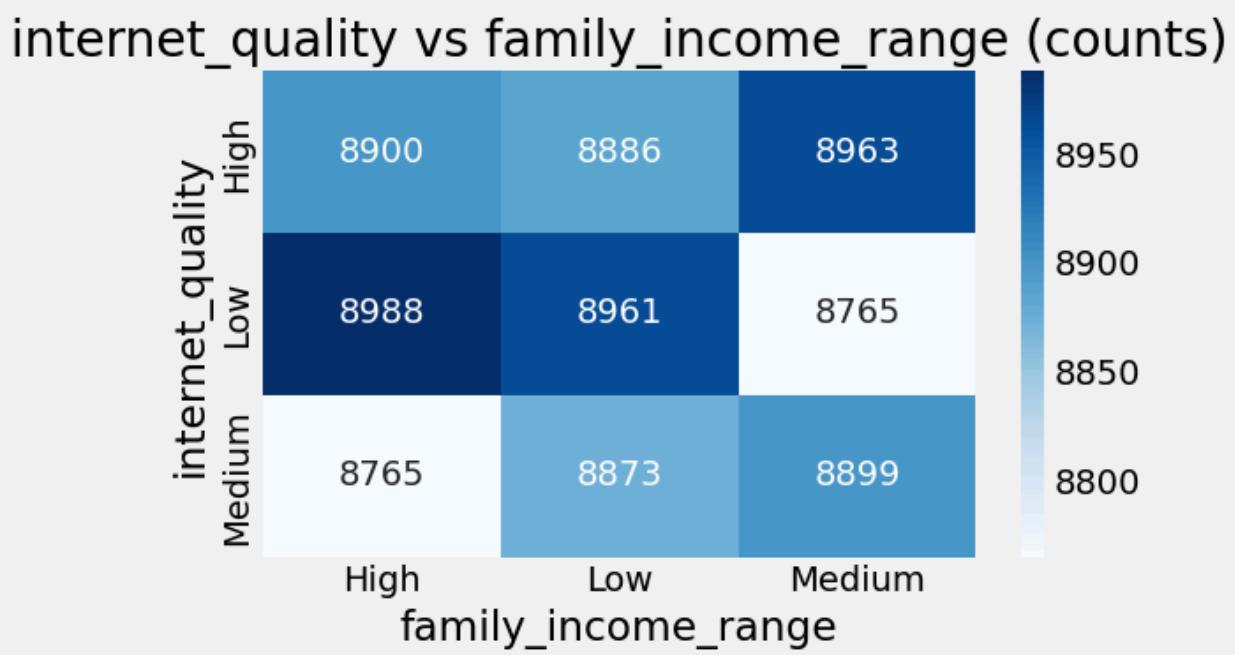
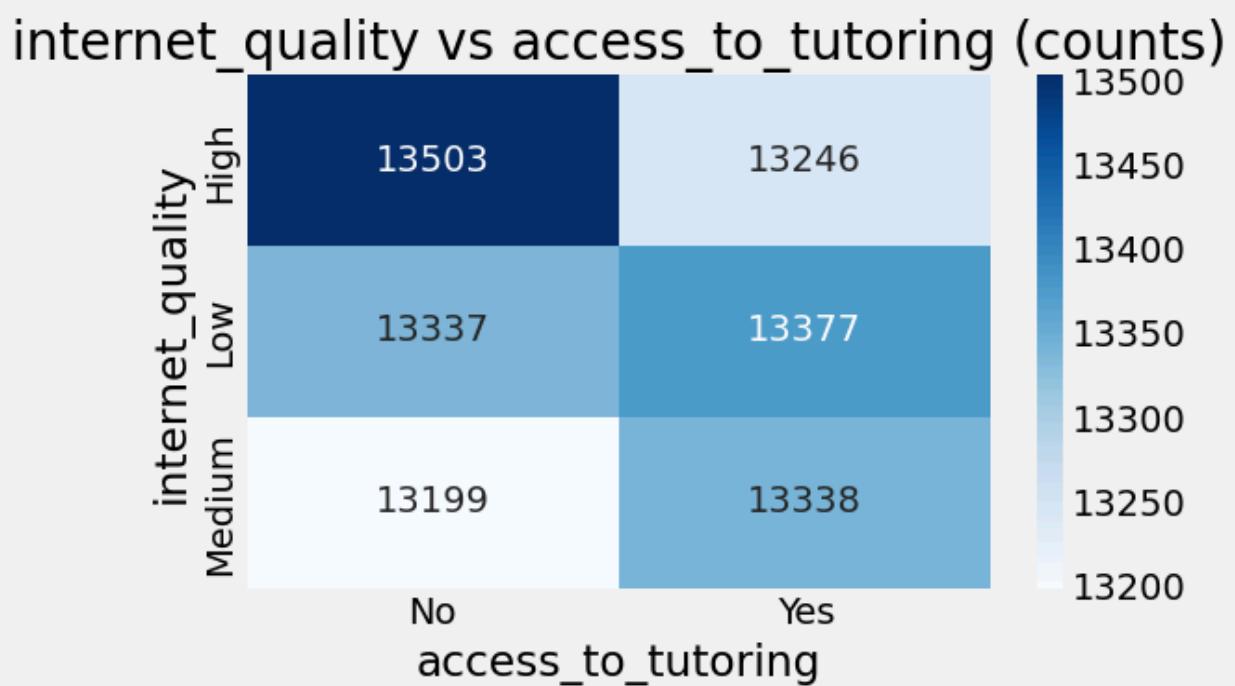


internet_quality vs dropout_risk (counts)

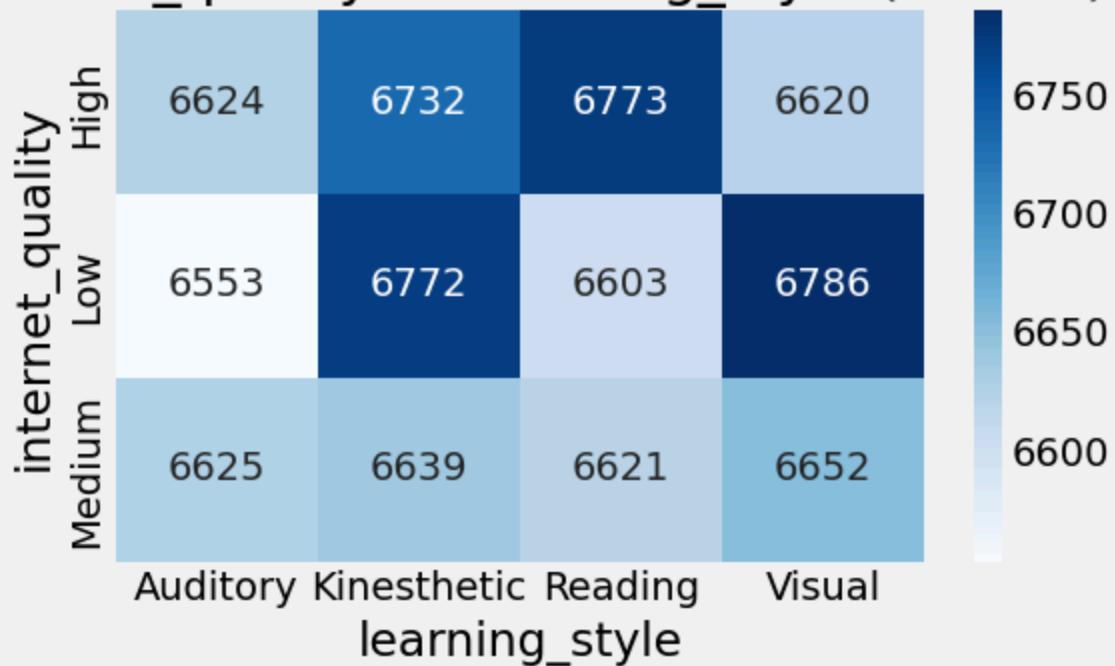


internet_quality vs study_environment (counts)

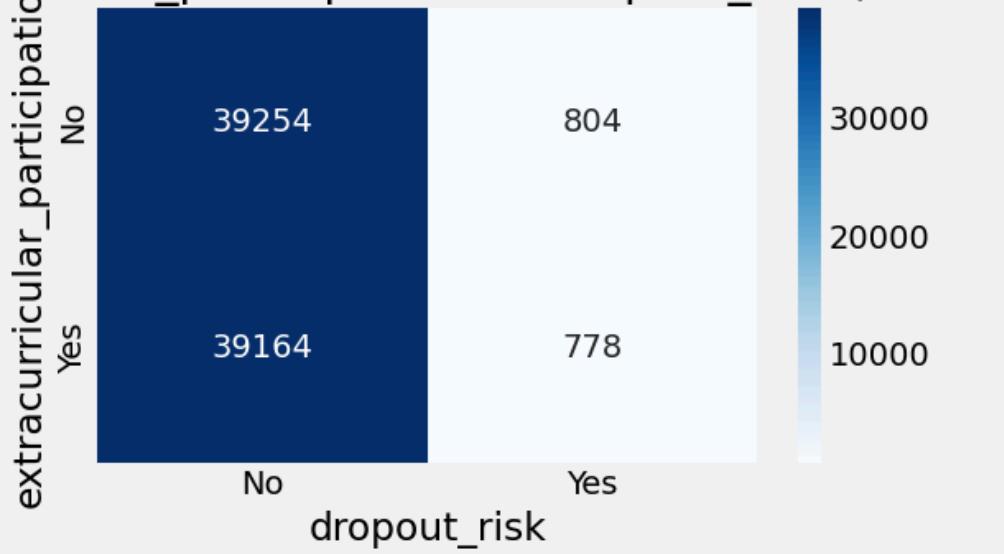


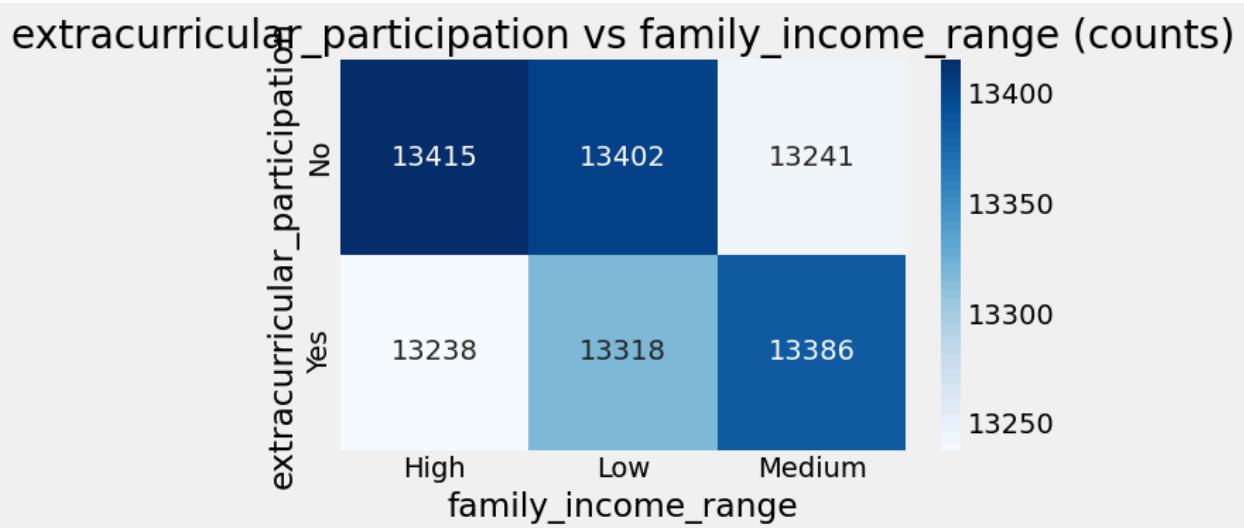
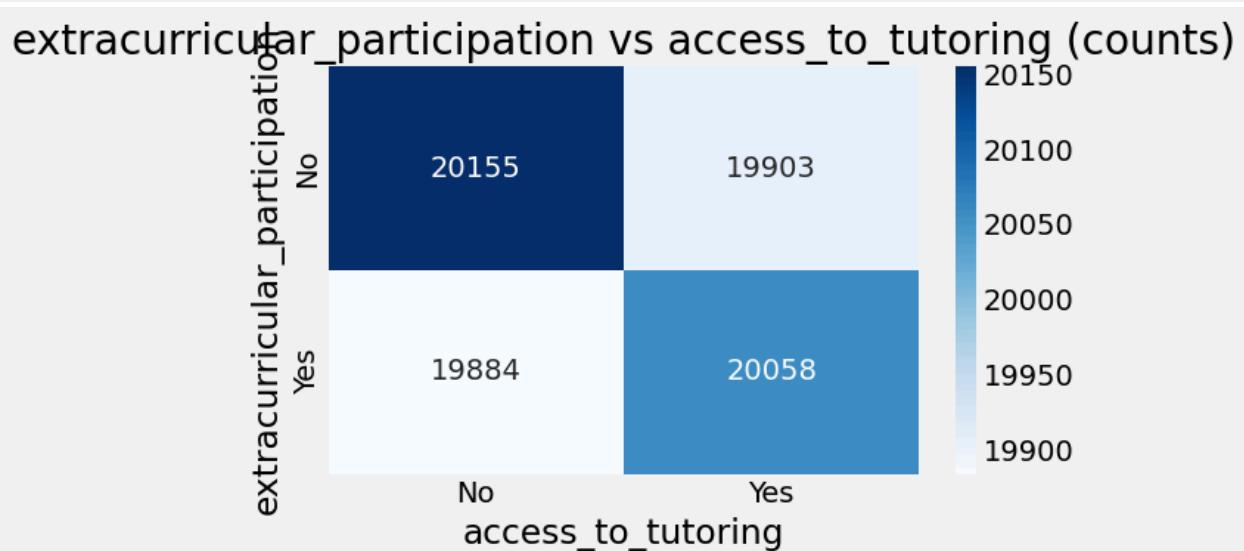
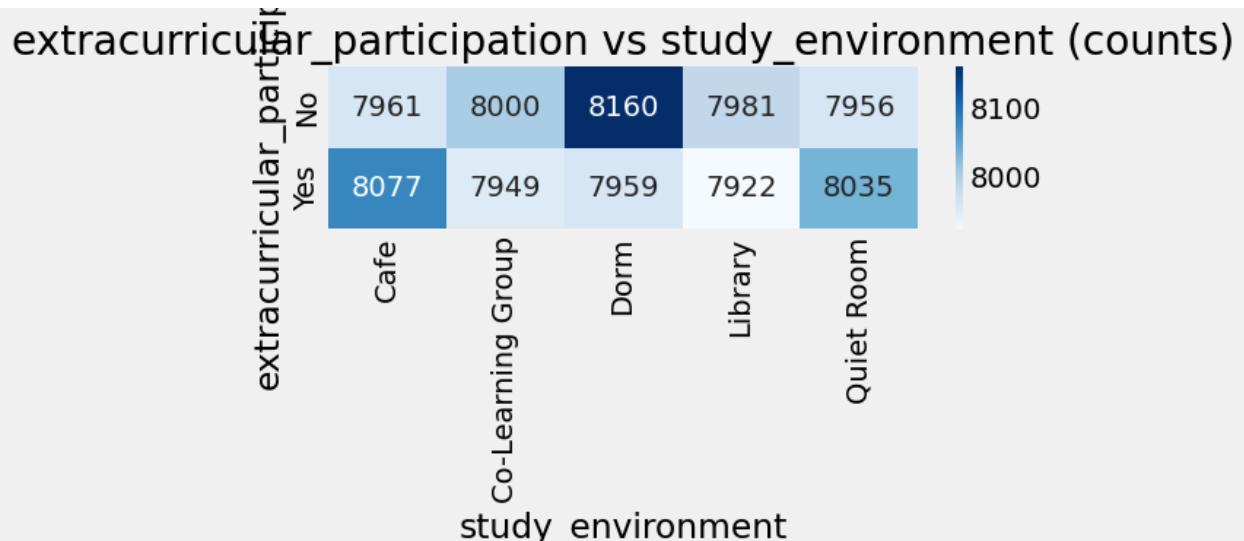


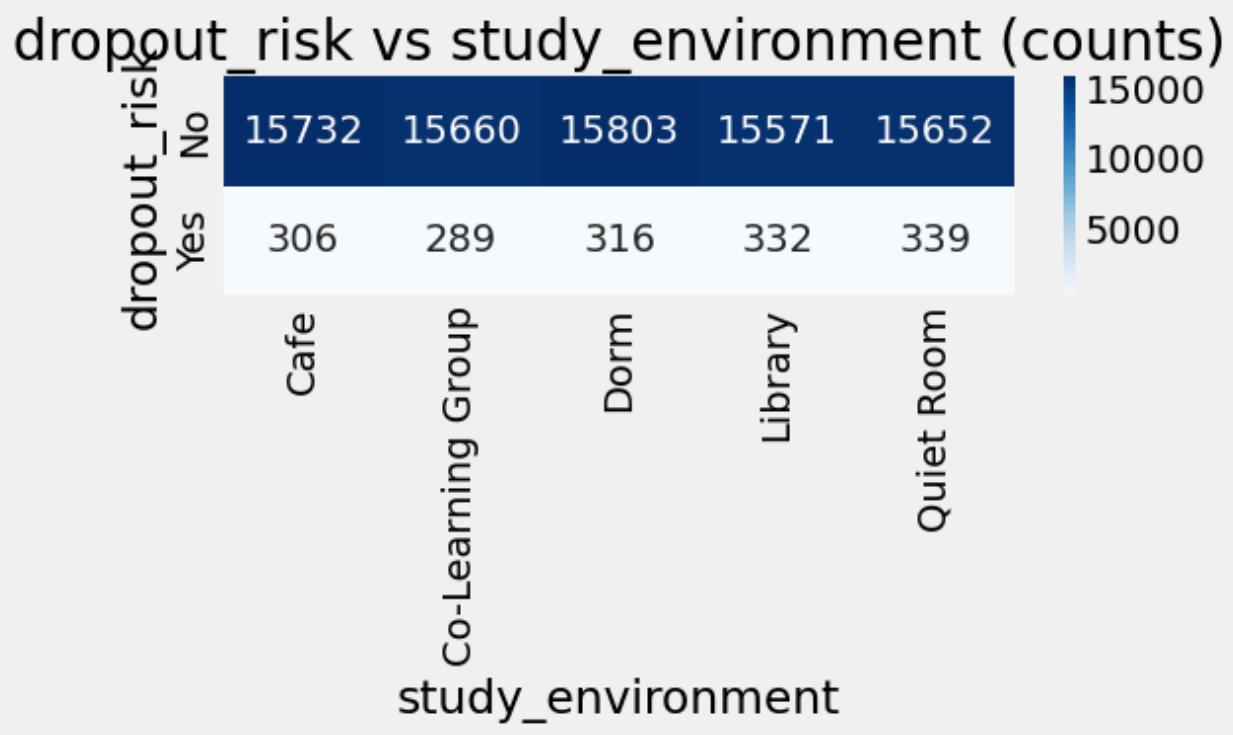
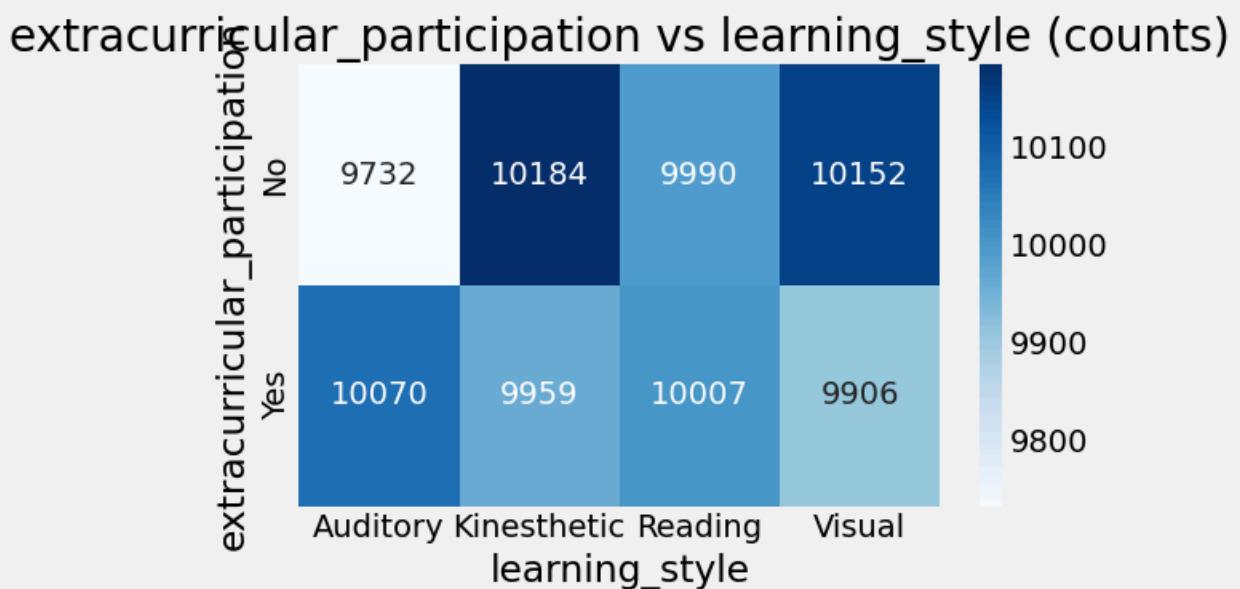
internet_quality vs learning_style (counts)



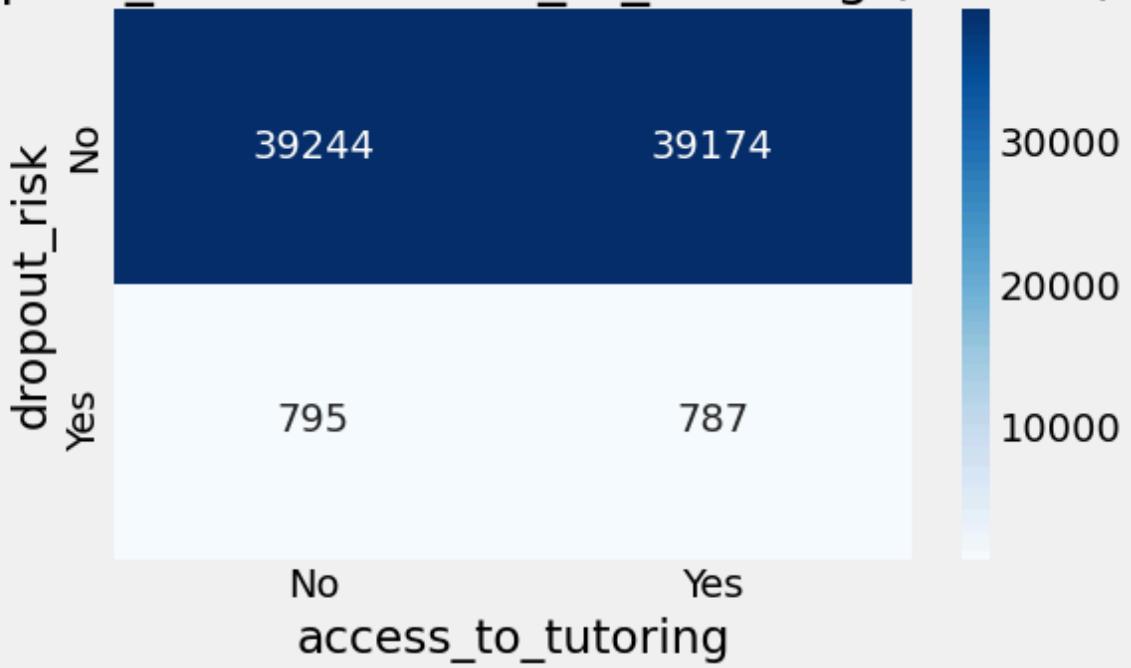
extracurricular_participation vs dropout_risk (counts)



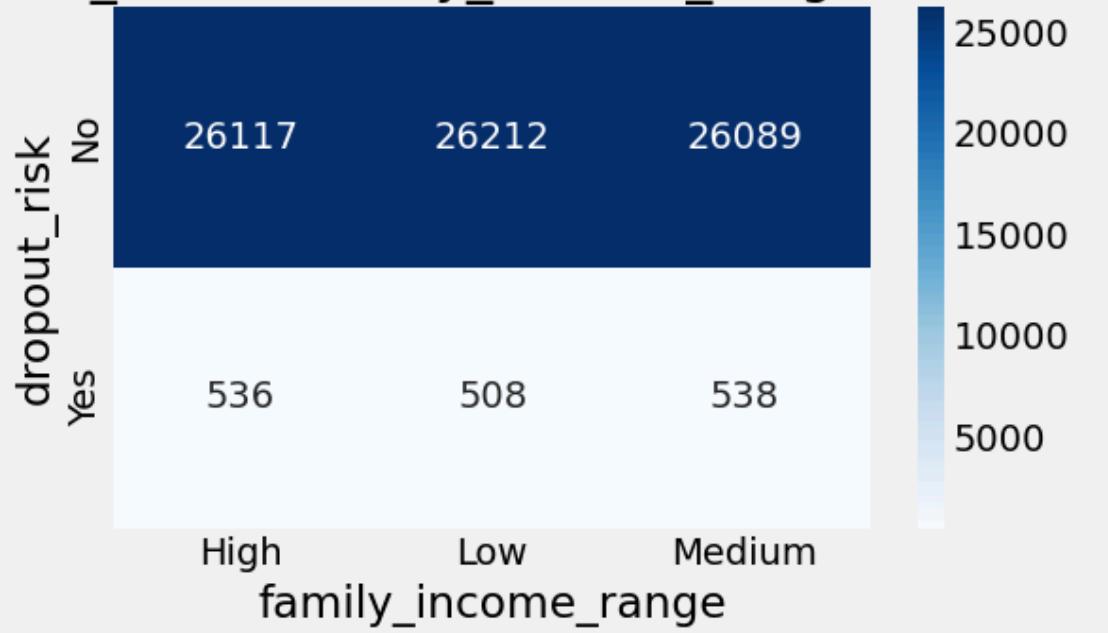




dropout_risk vs access_to_tutoring (counts)



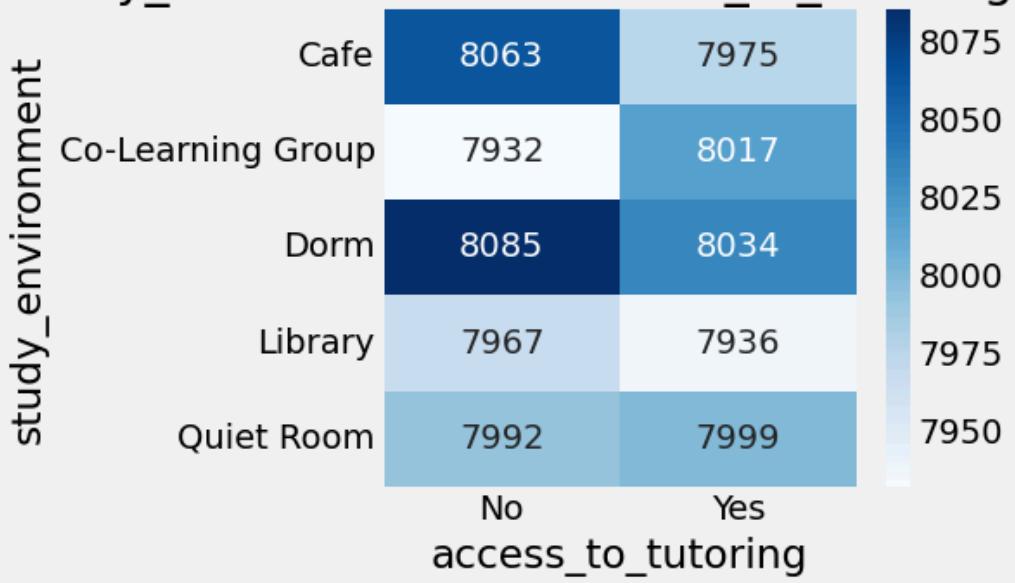
dropout_risk vs family_income_range (counts)



dropout_risk vs learning_style (counts)



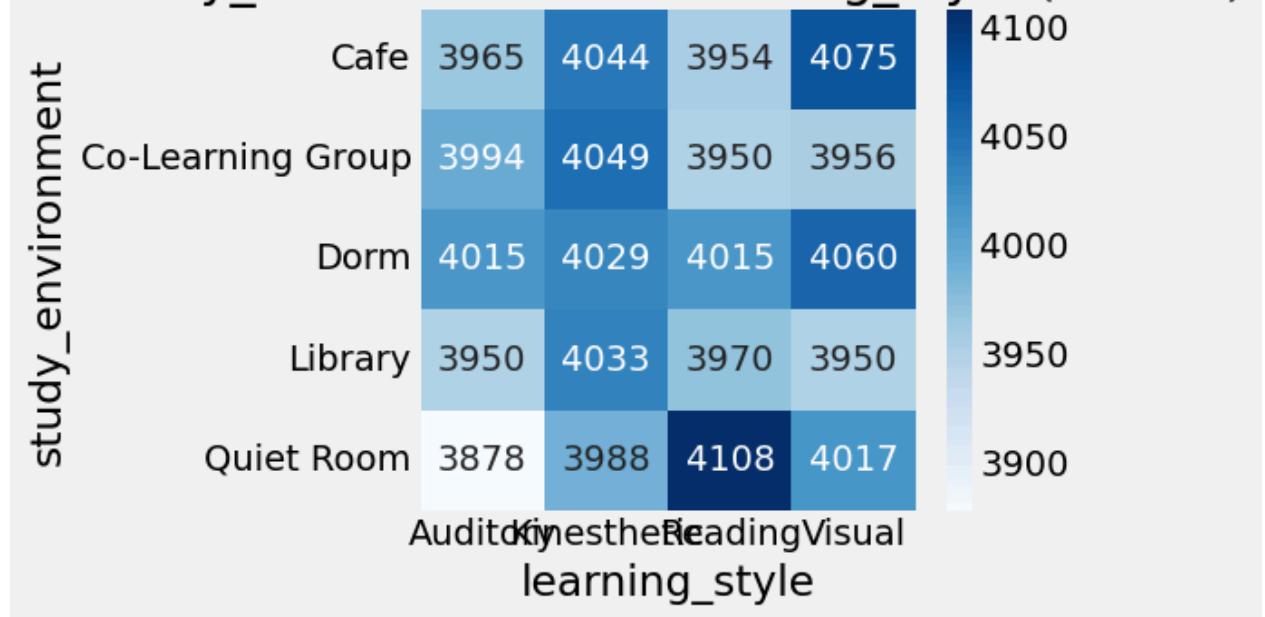
study_environment vs access_to_tutoring (counts)

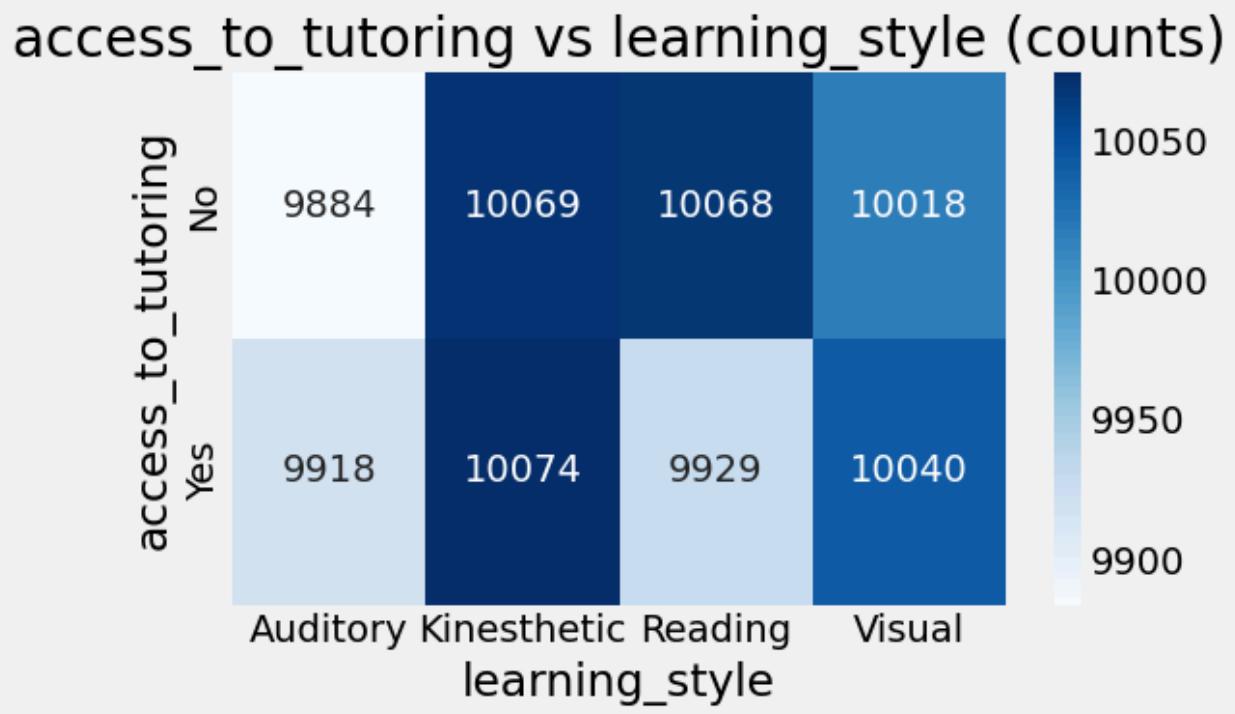
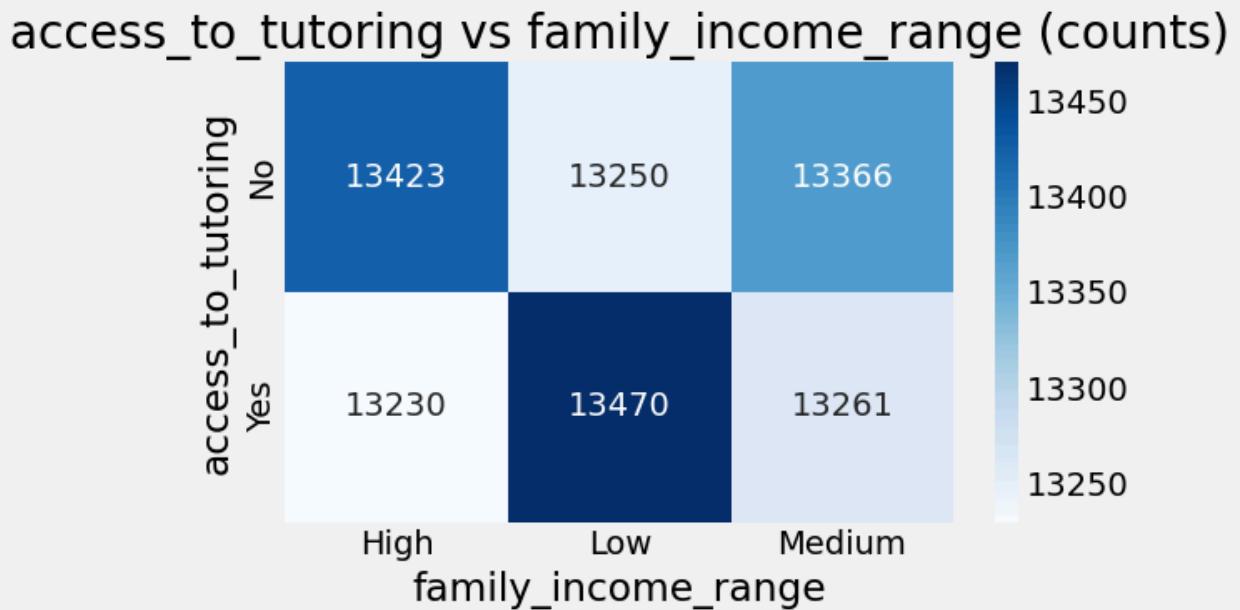


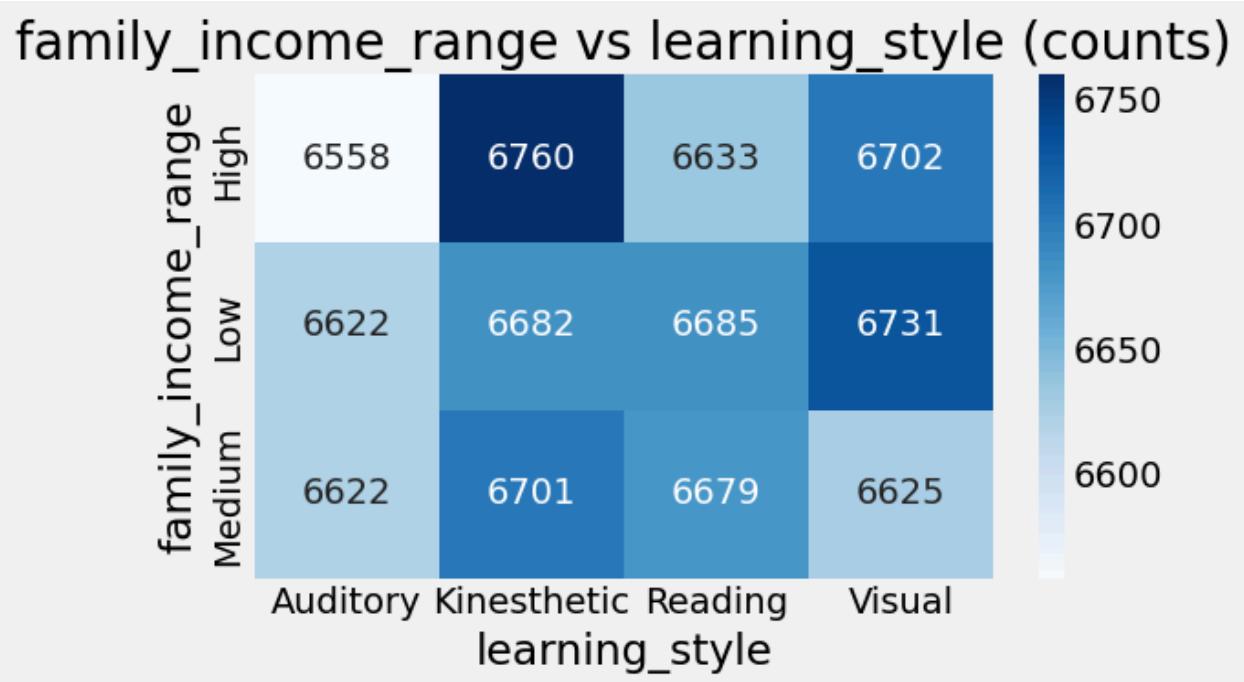
study_environment vs family_income_range (counts)



study_environment vs learning_style (counts)







```
In [ ]: students.sort("previous_gpa").select("previous_gpa", "time_management_score").show(5)
```

```
In [14]: ed_level = students["parental_education_level"]
ed_level
```

```
Out[14]: 0      Some College
1                  PhD
2      High School
3          Master
4          PhD
...
79995    High School
79996    Some College
79997        Master
79998          PhD
79999    Some College
Name: parental_education_level, Length: 80000, dtype: object
```

```
In [ ]: good_gpa=students.where("previous_gpa",are.above_or_equal_to(3.5))
good_gpa
```

```
In [ ]: pass_exam=students.where("exam_score",are.above_or_equal_to(90))
pass_exam
```

```
In [ ]: def stem(x):
        if x=="Psychology":
            return "Nonstem"
        if x=="Biology":
            return "Stem"
        if x=="Engineering":
            return "Stem"
```

```
if x=="Computer Science":  
    return "Stem"  
if x=="Arts":  
    return "Nonstem"  
if x=="Business":  
    return "Nonstem"
```

```
In [ ]: stem_nonstem=students.apply(stem,"major")
```

```
In [ ]: stem_categorized=students.with_column("Stem/Nonstem",stem_nonstem)
```

```
In [ ]: stem_only=stem_categorized.where("Stem/Nonstem","Stem")
```

```
In [ ]: pass_stem=stem_only.where("exam_score",are.above_or_equal_to(75))
```

```
In [ ]: np.average(pass_stem.column("study_hours_per_day"))
```

```
In [ ]: np.average(pass_stem.column("social_media_hours"))
```

```
In [ ]: np.average(fail_stem.column("study_hours_per_day"))
```

```
In [ ]: np.average(fail_stem.column("social_media_hours"))
```

```
In [ ]: fail_stem=stem_only.where("exam_score",are.below(75))
```

```
In [ ]: nonstem_only=stem_categorized.where("Stem/Nonstem","Nonstem")
```

```
In [ ]: nonstem_only.groupby("dropout_risk")
```

```
In [ ]: np.average(pass_exam.column("time_management_score"))
```

```
In [ ]: np.average(fail_exam.column("time_management_score"))
```

```
In [ ]: fail_exam=students.where("exam_score",are.below(75))  
fail_exam
```

```
In [ ]: np.average(good_gpa.column("time_management_score"))
```

```
In [ ]: np.average(good_gpa.column("sleep_hours"))
```

```
In [ ]: np.average(bad_gpa.column("sleep_hours"))
```

```
In [ ]: bad_gpa=students.where("previous_gpa",are.below(2))  
bad_gpa
```

```
In [ ]: np.average(bad_gpa.column("study_hours_per_day"))
```

```
In [ ]: Table.read_table("CollegeScores4yr.csv")  
  
In [ ]: bad_gpa.select("motivation_level").group("motivation_level")  
  
In [ ]: good_gpa.select("motivation_level").group("motivation_level")  
  
In [ ]: time_manage=students.select("time_management_score","previous_gpa").sort("time  
In [ ]: factors_affecting=Table.read_table("Factors_ affecting_ university_student_gra  
In [ ]: factors_affecting.group("Motivation")  
  
In [ ]: study_habits=Table.read_table("student_study_habits.csv")  
study_habits  
  
In [ ]: stem_2019=Table.read_table("stem_2019.csv")  
stem_2019  
  
In [ ]: stem_emp_2019=stem_2019.select("STEM_group","tot_emp").take(0).relabeled("tot_  
stem_emp_2019  
  
In [ ]: stem_2020=Table.read_table("stem_2020.csv")  
stem_emp_2020=stem_2020.select("STEM_group","tot_emp").take(0).relabeled("tot_  
stem_emp_2020  
  
In [ ]: two_years=stem_emp_2019.join("STEM_group",stem_emp_2020,"STEM_group")  
  
In [ ]: stem_emp_table=Table().with_columns("Year",make_array(2019,2020,2021,2022,2023  
"STEM Employment Total",make_array(9345230,9321470,926319  
In [ ]: stem_emp_years=stem_emp_table.column("Year")  
stem_emp_tot=stem_emp_table.column("STEM Employment Total")  
  
plt.plot(stem_emp_years,stem_emp_tot)  
plt.title('STEM Employment Over the Years')  
plt.xlabel('Year')  
plt.ylabel('STEM Employment Total')  
plt.ylim(9200000, 10300000) # optional: set y-axis range # shows the labels  
plt.grid(True)  
  
In [ ]:  
  
In [ ]: stem_2021=Table.read_table("stem_2021.csv")  
stem_emp_2021=stem_2021.select("STEM_group","tot_emp").take(0)  
stem_emp_2021  
  
In [ ]: stem_2022=Table.read_table("stem_2022.csv")  
stem_emp_2022=stem_2022.select("STEM_group","tot_emp").take(0)  
stem_emp_2022
```

```
In [ ]: stem_2023=Table.read_table("stem_2023.csv")
stemp_emp_2023=stem_2023.select("STEM_group","tot_emp").take(0)
stemp_emp_2023
```

```
In [ ]: stem_2024=Table.read_table("stem_2024.csv")
stemp_emp_2024=stem_2024.select("STEM_group","tot_emp").take(0)
stemp_emp_2024
```

```
In [ ]: stem_degrees=Table.read_table("tabn318.45.csv")
stem_degrees
```

```
In [ ]: stem_degrees.take(np.arange(6,16)).select(0,1) #.relabeled("Table 318.45. Number of STEM Degrees Granted by Race/Ethnicity in 2012-2022")
```

```
In [ ]: privileged=stem_degrees.take(np.arange(6,16)).select(1,2)
```

```
In [ ]: privileged
```

```
In [ ]: privileged_table=Table().with_columns("Total", make_array(556696,574000,604167),
                                             "White",make_array(333652,337241,348586,
                                             privileged_table.column(1)/privileged_table.column(0))
```

```
In [ ]: stem_degree_table=Table().with_columns("Year",np.arange(2012,2022),
                                              "Total STEM Degrees", make_array(556696,574000,604167))
stem_degree_table
```

```
In [ ]: stem_degree_years=stem_degree_table.column("Year")
stem_degree_granted=stem_degree_table.column("Total STEM Degrees")

plt.plot(stem_degree_years,stem_degree_granted)
plt.title('STEM Degrees Granted Over the Years')
plt.xlabel('Year')
plt.ylabel('Total STEM Degrees')
plt.ylim(500000, 800000) # optional: set y-axis range    # shows the labels
plt.grid(True)
```

```
In [ ]: Table.read_table("education_2024.csv")
```

```
In [16]: women=Table.read_table("women_in_stem.csv")
women
#female_enrollment=women.sort("Year").select("Year","Female Enrollment (%)").group_by("Year").sum("Female Enrollment (%)")
```

Out[16]:

Country	Year	Female Enrollment (%)	Female Graduation Rate (%)	STEM Fields	Gender Gap Index
China	2018	20.4	43.2	Engineering	0.52
China	2005	35.6	29.3	Mathematics	0.98
China	2005	53.7	32.4	Biology	0.6
Germany	2007	65	63.6	Mathematics	0.69
Canada	2010	54.4	28.8	Engineering	0.74
India	2001	53	54.5	Computer Science	0.54
Canada	2002	69	23.2	Engineering	0.86
India	2000	29.5	38.1	Engineering	0.9
USA	2014	57.1	35.7	Biology	0.61
China	2020	36.7	64.3	Computer Science	0.96

... (490 rows omitted)

```
In [ ]: women.where("Country", "Australia").select("Year", "Female Enrollment (%)").groupby("Year").mean().show()
```

```
In [ ]: india_gender_gap=women.where("Country", "India").sort("Year").select("Year", "Gender Gap Index average")
india_gender_gap.plot("Year", "Gender Gap Index average")
```

```
In [ ]: print(india_grad)
print(us_grad)
```

```
In [17]: years=women.where("Country", "USA").sort("Year").select("Year", "Female Graduation Rate (%)")
india_grad=women.where("Country", "India").sort("Year").select("Year", "Female Graduation Rate (%)")
us_grad=women.where("Country", "USA").sort("Year").select("Year", "Female Graduation Rate (%)")
australia_grad=women.where("Country", "Australia").select("Year", "Female Graduation Rate (%)")
china_grad=women.where("Country", "China").sort("Year").select("Year", "Female Graduation Rate (%)")
canada_grad=women.where("Country", "Canada").sort("Year").select("Year", "Female Graduation Rate (%)")
germany_grad=women.where("Country", "Germany").sort("Year").select("Year", "Female Graduation Rate (%)")

plt.plot(india_years, india_grad, label="India", marker="o")
plt.plot(india_years, australia_grad, label="Australia")
# plt.plot(years, australia_grad, label="Australia")
# plt.plot(years, china_grad, label="China")
# plt.plot(years, canada_grad, label="Canada")
# plt.plot(years, germany_grad, label="Germany")

plt.title('Female STEM Graduation Rates')
plt.xlabel('Year')
plt.ylabel('Percentage')
plt.ylim(0, 60) # optional: set y-axis range
plt.legend() # shows the labels
```

```
plt.grid(True)

-----
NameError                                 Traceback (most recent call last)
Cell In[17], line 9
      6 canada_grad=women.where("Country","Canada").sort("Year").select("Year","Female Graduation Rate (%)").group("Year",np.average).column("Female Graduation Rate (%) average")
      7 germany_grad=women.where("Country","Germany").sort("Year").select("Year","Female Graduation Rate (%)").group("Year",np.average).column("Female Graduation Rate (%) average")
----> 9 plt.plot(india_years,india_grad,label="India",marker="o")
     10 plt.plot(india_years,australia_grad,label="USA")
     11 #plt.plot(years,australia_grad,label="Australia")
     12 #plt.plot(years,china_grad,label="China")
     13 #plt.plot(years,canada_grad,label="Canada")
     14 #plt.plot(years,germany_grad,label="Germany")

NameError: name 'india_years' is not defined
```

```
In [ ]: change_majors=Table.read_table("pre_and_post_semester_majors.csv")
stem_pre=change_majors.where("premaj","stem")
stem_pre.select("premaj","pstmaj").group("pstmaj")

In [ ]: grad_students=Table.read_table("graduated_students.csv")
first_gen=grad_students.where("generation","first").where("persist",are.not_equal)
no_first=grad_students.where("generation","notfst").where("persist",are.not_equal)

first_gen.bahr("persist","count")

In [ ]: categories=make_array("Stayed STEM", "Left STEM")
first_genners=make_array(76,24)
non_firsty=make_array(561,70)

x = np.arange(len(categories)) # positions for groups
width = 0.35 # bar width

fig, ax = plt.subplots(figsize=(8,6))

bars1 = ax.bar(x - width/2, first_genners, width, label='First-Gen', color="#f9a86a")
bars2 = ax.bar(x + width/2, non_firsty, width, label='Non-First-Gen', color="#6a99cc")

ax.set_ylabel('Number of Students')
ax.set_title('STEM Retention: First-Gen vs. Non-First-Gen')
ax.set_xticks(x)
ax.set_xticklabels(categories)
ax.legend()

for bars in [bars1, bars2]:
    for bar in bars:
        height = bar.get_height()
```

```
        ax.text(bar.get_x() + bar.get_width()/2., height + 5,
                 f'{height}', ha='center', va='bottom')
```

```
In [ ]: categories_gen=make_array("First Gen", "Not First Gen")
first_genners=make_array(76,24)
non_firsty=make_array(561,70)

x = np.arange(len(categories)) # positions for groups
width = 0.35 # bar width

fig, ax = plt.subplots(figsize=(8,6))

bars1 = ax.bar(x - width/2, first_genners, width, label='First-Gen', color='#f9a86a')
bars2 = ax.bar(x + width/2, non_firsty, width, label='Non-First-Gen', color='#6a9af9')

ax.set_ylabel('Number of Students')
ax.set_title('STEM Retention: First-Gen vs. Non-First-Gen')
ax.set_xticks(x)
ax.set_xticklabels(categories)
ax.legend()

for bars in [bars1, bars2]:
    for bar in bars:
        height = bar.get_height()
        ax.text(bar.get_x() + bar.get_width()/2., height + 5,
                f'{height}', ha='center', va='bottom')
```

```
In [ ]: grad_students.group("generation")
```

```
In [ ]: grad_students.where("generation","notfst").group("persist")
```

```
In [ ]: us_grad_avg=np.average(women.where("Country","USA")).column("Female Graduation Rate (%)")
india_grad_avg=np.average(women.where("Country","India")).column("Female Graduation Rate (%)")
china_grad_avg=np.average(women.where("Country","China")).column("Female Graduation Rate (%)")
germany_grad_avg=np.average(women.where("Country","Germany")).column("Female Graduation Rate (%)")
canada_grad_avg=np.average(women.where("Country","Canada")).column("Female Graduation Rate (%)")
australia_grad_avg=np.average(women.where("Country","Australia")).column("Female Graduation Rate (%)")

bargraph_table=Table().with_columns("Country",make_array("China","USA","Germany","India","Australia","Canada"),
                                         "Average Female STEM Graduation Rate (%)",
                                         "Average Female STEM Enrollment (%)")
bargraph_table.barch("Country","Average Female STEM Graduation Rate (%)")
```

```
In [ ]: us_enrollment=women.where("Country","USA").sort("Year").select("Year","Female Enrollment (%)")
india_enrollment=women.where("Country","India").sort("Year").select("Year","Female Enrollment (%)")
india_years=women.where("Country","India").sort("Year").select("Year","Female Enrollment (%)")
india_grad=women.where("Country","India").sort("Year").select("Year","Female Graduation Rate (%)")

plt.plot(india_years,india_grad,label="USA graduation rate")
plt.plot(india_years,india_enrollment,label="USA enrollment")
```

```
plt.title('Female STEM Graduation Rates vs Enrollment')
plt.xlabel('Year')
plt.ylabel('Percentage')
plt.ylim(0, 60) # optional: set y-axis range
plt.legend()      # shows the labels
plt.grid(True)
```

```
In [ ]: india_years
```

```
In [ ]: india_graduation=women.where("Country","India").sort("Year").select("Year","Female_Graduation_Rate_(%) average")
```

```
In [ ]: us_gender_gap=women.where("Country","USA").sort("Year").select("Year","Gender Gap Index average")
```

```
In [ ]: female_enrollment.plot("Year","Female Enrollment (%) average")
```

```
In [ ]: us_enroll=women.where("Country","USA").sort("Year").select("Year","Female Enrollment (%) average")
```

```
In [ ]: us_enroll.plot("Year","Female Enrollment (%) average")
```

```
In [ ]: us_grad=women.where("Country","USA").sort("Year").select("Year","Female Graduation Rate (%) average")
```

```
In [ ]: australia_grad=women.where("Country","Australia").sort("Year").select("Year","Female Graduation Rate (%) average")
```

```
In [ ]: india_enroll=women.where("Country","India").sort("Year").select("Year","Female Enrollment (%) average")
```

```
In [ ]: india_enroll.plot("Year","Female Enrollment (%) average")
```

```
In [ ]: female_grad=women.sort("Year").select("Year","Female Graduation Rate (%) average")
```

```
In [ ]: india_women=women.where("Country","India").sort("Year").select("Year","Female Graduation Rate (%) average")
```

```
In [ ]: india_women.plot("Year","Female Graduation Rate (%) average")
```

```
In [ ]: australia_grad=women.where("Country","Australia").select("Year","Female Graduation Rate (%) average")
```

```
In [ ]: all_stem=Table.read_table("UNESCO_UIS_GRAD_STEM.csv")
india_all=all_stem.where("REF_AREA_LABEL",are.equal_to("India")).sort("TIME_PERIOD")
india_all
```

```
In [ ]: usa_all=all_stem.where("REF_AREA_LABEL",are.equal_to("United States")).sort("TIME_PERIOD")
usa_all
```

```
In [ ]: all_stem.where("REF_AREA_LABEL","United States").where("SEX","_T")
```

```
In [ ]: student_data=Table.read_table("student_performance_dataset.csv")
student_data
```

```
In [ ]: student_data.where("Final_Exam_Score",50)
```

```
In [ ]: student_data.sort("Study_Hours_per_Week").select("Study_Hours_per_Week","Final
```

```
In [ ]: Table.read_table("ResearchInformation3.csv").group("Department")
```

```
In [ ]:
```