Tecnológico de Monterrey

Campus Ciudad de México

Activity 4: Heatmaps and boxplots

Ximena Silva Bárcena

A01785518

Rodrigo Martínez Vallejo

A00573055

Prof. Sergio Ruiz Loza

Mastering Analytics

Grupo **201**

Wednesday, May 7th, 2025

Ximena Silva Bárcena
Are there any variables that do not provide information?
Variables like parental_education_level and internet_quality show very low correlation with the exam_score and other key performance indicators. This suggests they might not significantly influence the overall academic performance of student
If you had to eliminate variables, which ones would you remove and why?
Parental_education_level and internet_quality might also be candidates for removal if their correlation with academic performance is weak or if they introduce unnecessary complexity without significantly contributing to the model's predictive power.
Are there any variables with unusual data?
The mental_health_rating variable shows a wide range with extreme low and high values, which might indicate potential outliers or inconsistent data entries. The attendance_percentage also appears to have values close to 100, suggesting it might be skewed or have limited variance.
If you compare the variables, are they all in similar ranges?
The variables have different scales. For example, sleep_hours ranges from 3.2 to 10, while attendance_percentage ranges from 0 to 100. This wide variation in scales can distort distance-based algorithms like K-Means, leading to misleading clustering results if not properly normalized.
Do you think this affects the data analysis? Can you find any similar groups? What are these groups?
Using unscaled data can lead to misleading results in cluster analysis because features with larger numerical ranges dominate the distance calculations. For example, early cluster tests might incorrectly separate students based on attendance rates or sleep hours alone, ignoring other critical factors like study

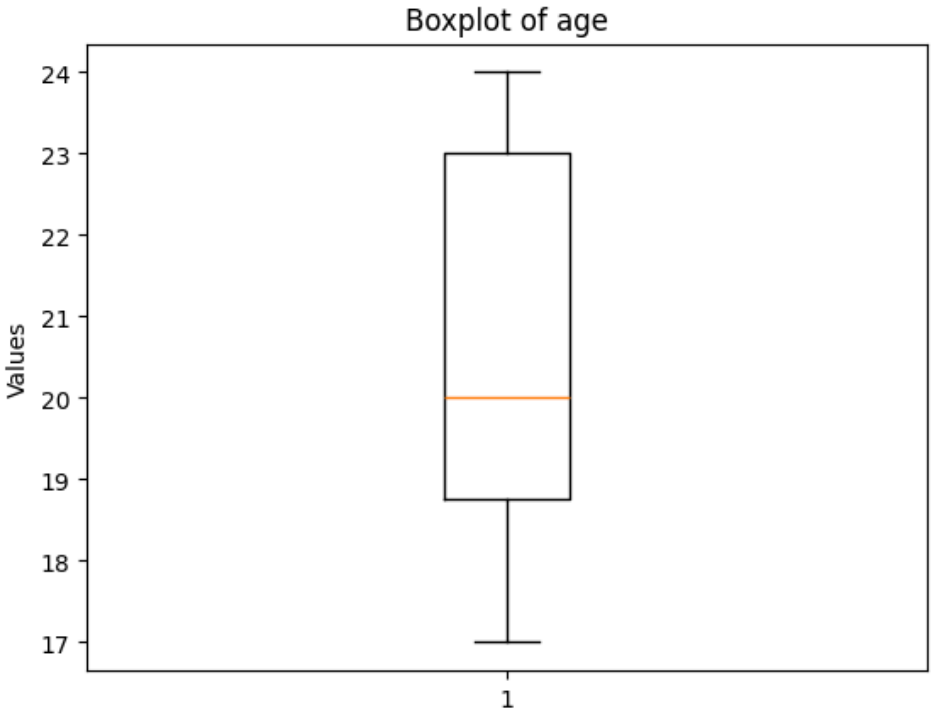
habits or mental health. To avoid this, it's crucial to normalize the data, revealing more accurate groupings, like diligent students with high exam scores and balanced lifestyles, versus distracted students with high entertainment consumption and lower academic outcomes.

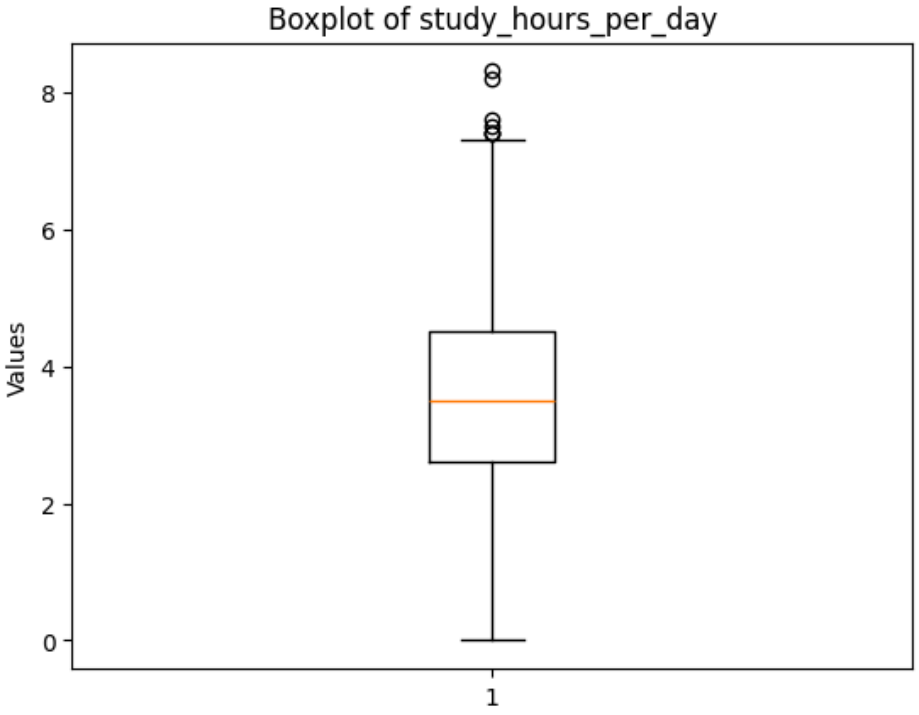
Rodrigo Martínez Vallejo
Are there any variables that do not provide information?
Parental Education Level and Internet Quality both show very weak correlations with exam scores, making them poor predictors of academic success. Similarly, Extracurricular Participation has almost no correlation with exam results, suggesting it doesn't contribute valuable information to the analysis.
If you had to eliminate variables, which ones would you remove and why?
I would consider removing Parental Education Level, Internet Quality, and Extracurricular Participation because their weak correlations with exam scores indicate that they add little to no predictive power. Removing these variables can reduce model complexity and improve accuracy.
Are there any variables with unusual data?
Netflix Hours and Social Media Hours have extreme maximum values that likely indicate outliers or data entry mistakes, as it's unlikely that students spend this much time daily without severe academic impact.
If you compare the variables, are they all in similar ranges?
No, the ranges vary significantly. For example, attendance_percentage ranges from 0 to 100, while exercise_frequency is capped at 7. This imbalance can distort machine learning algorithms that rely on distance calculations, potentially skewing clustering results.

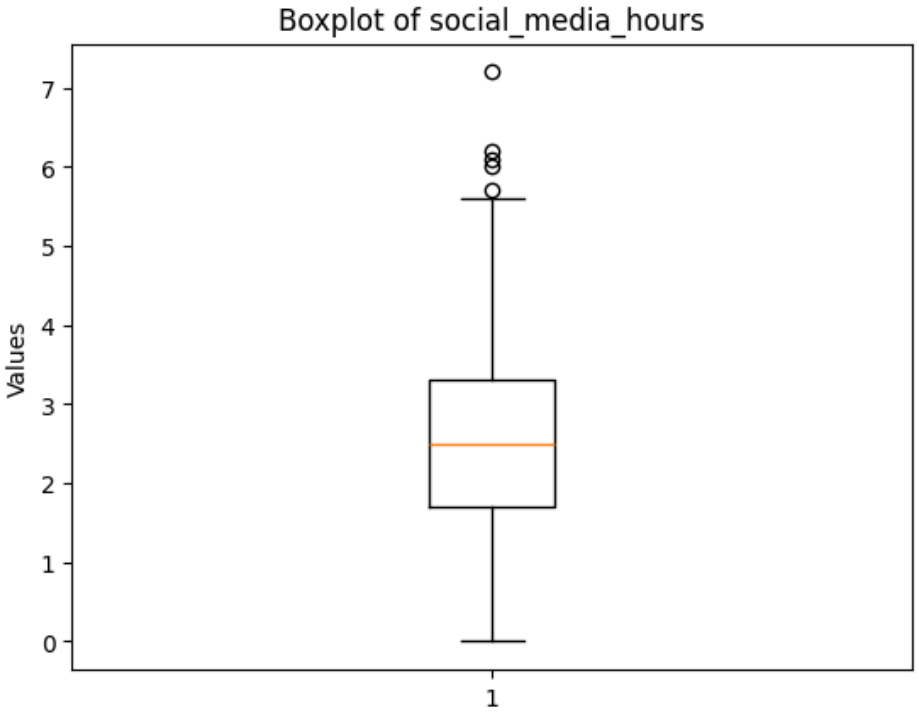
Do you think this affects the data analysis? Can you find any similar groups? What are these groups?
The differences in scales can affect data analysis by giving disproportionate weight to certain features. For instance, attendance percentages might overpower other metrics if not properly scaled. Preliminary clustering attempts suggest the existence of at least two main groups: students who study consistently with strong mental health and high scores, and another group characterized by excessive social media use, poor exam performance, and lower attendance.

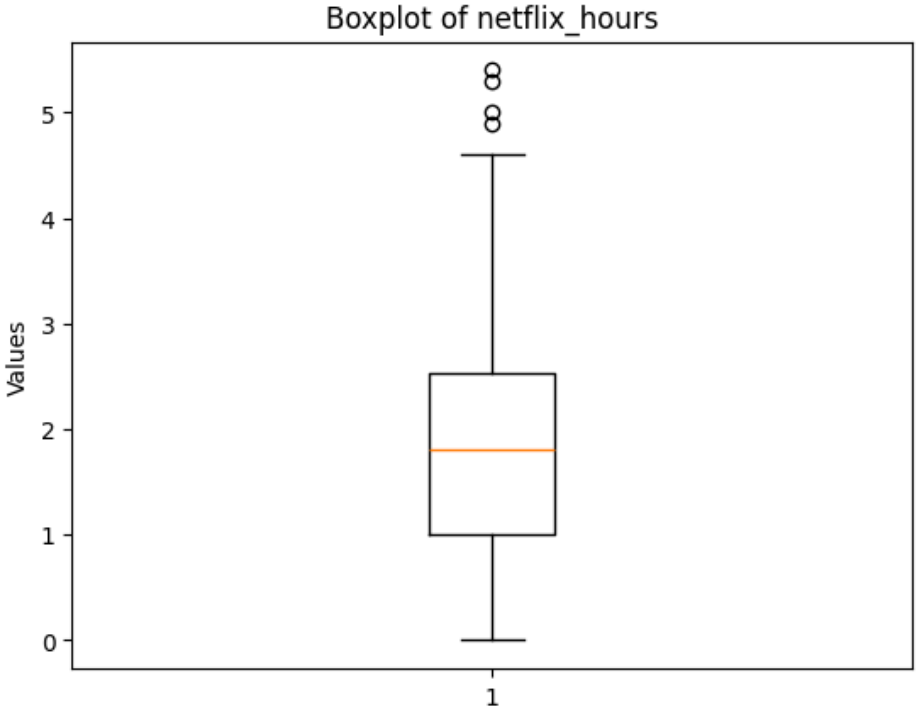
Box Diagram

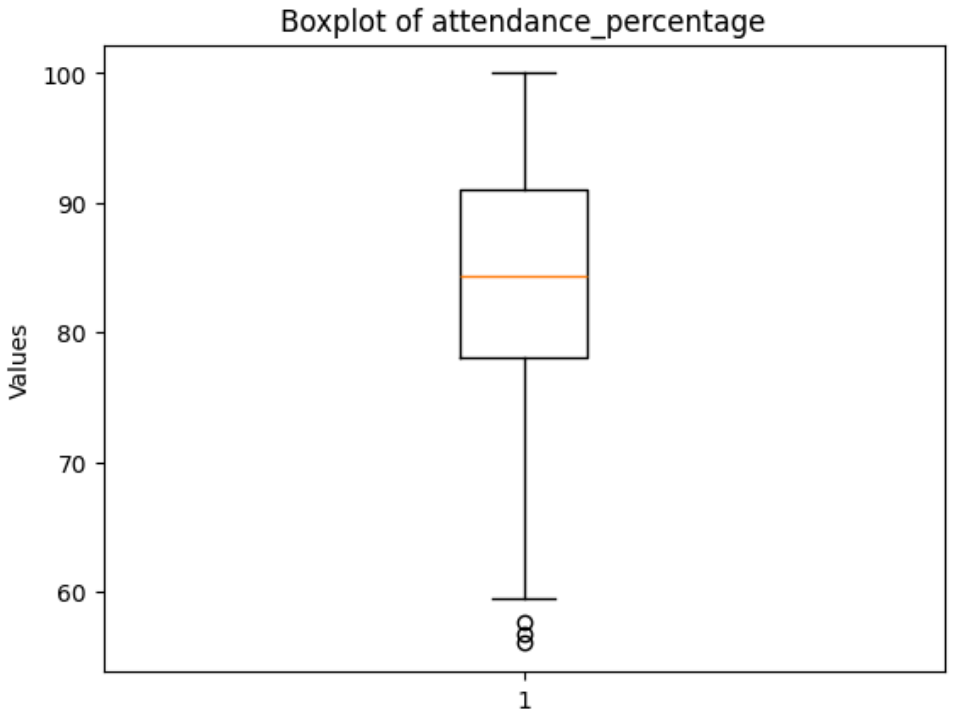
```
col_names = ['age', 'study_hours_per_day', 'social_media_hours', 'netflix_hours',
       'attendance_percentage', 'sleep_hours', 'exercise_frequency',
       'mental_health_rating', 'exam_score']

for name in col_names:
  plt.boxplot(numeric_data[name])
  plt.title(f'Boxplot of {name}')
  plt.ylabel('Values')
  plt.show()
```

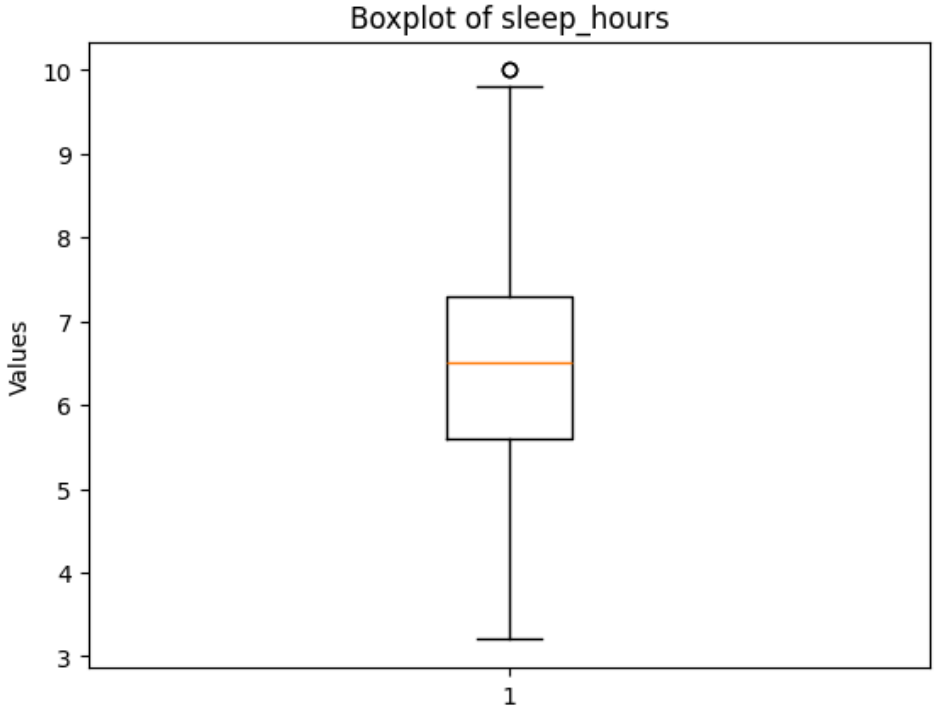

Boxplot of age
Values
24
23
22
21
20
19
18
17
1

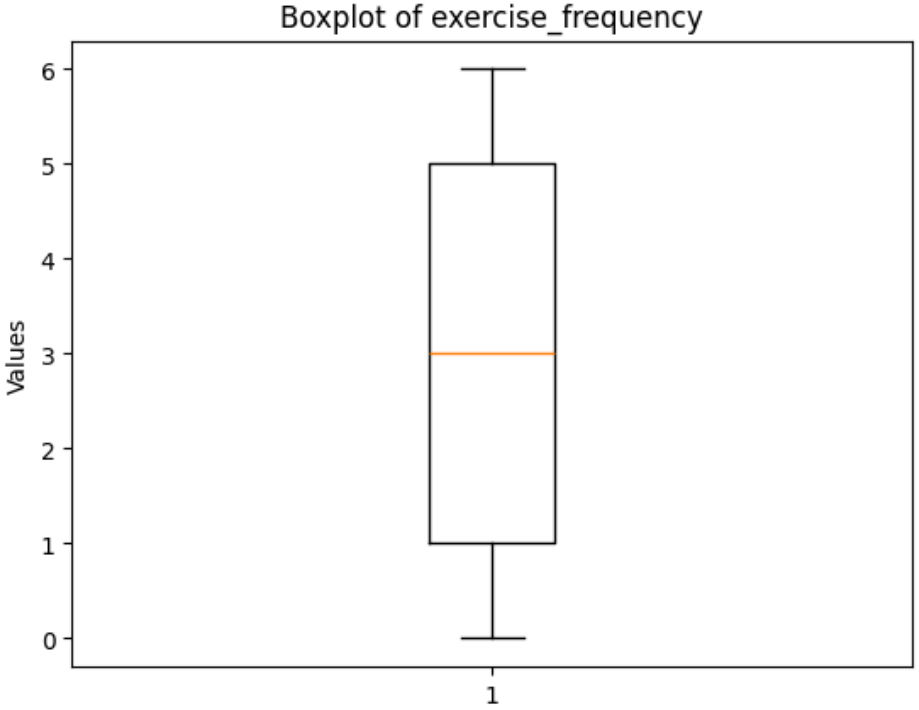

Boxplot of study_hours_per_day
Values
8
6
4
2
0
1

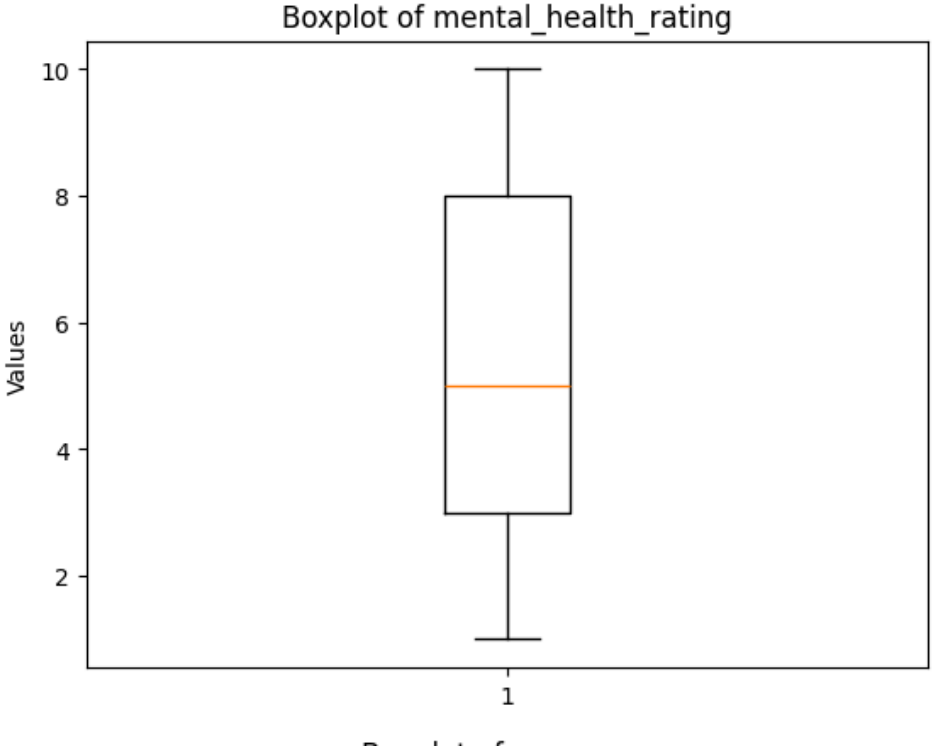

Boxplot of social_media_hours
Values
0
1
2
3
4
5
6
7
1

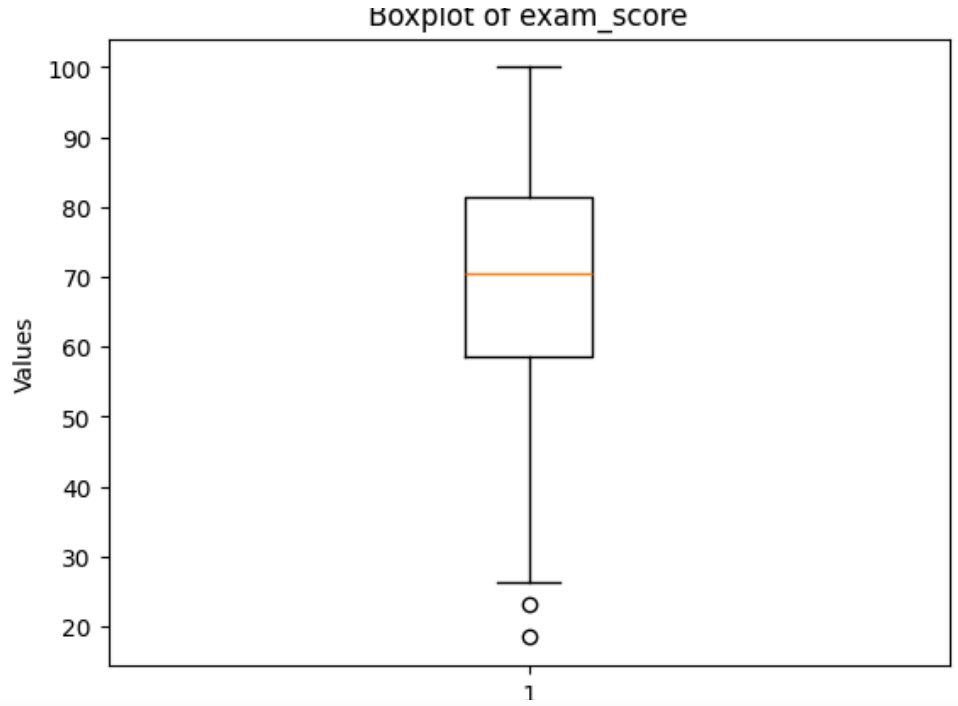

Boxplot of netflix_hours
Values
0
1
2
3
4
5
1


Boxplot of attendance_percentage
Values
100
90
80
70
60
1


Boxplot of sleep_hours
Values
10
9
8
7
6
5
4
3
1


Boxplot of exercise_frequency
Values
0
1
2
3
4
5
6
1

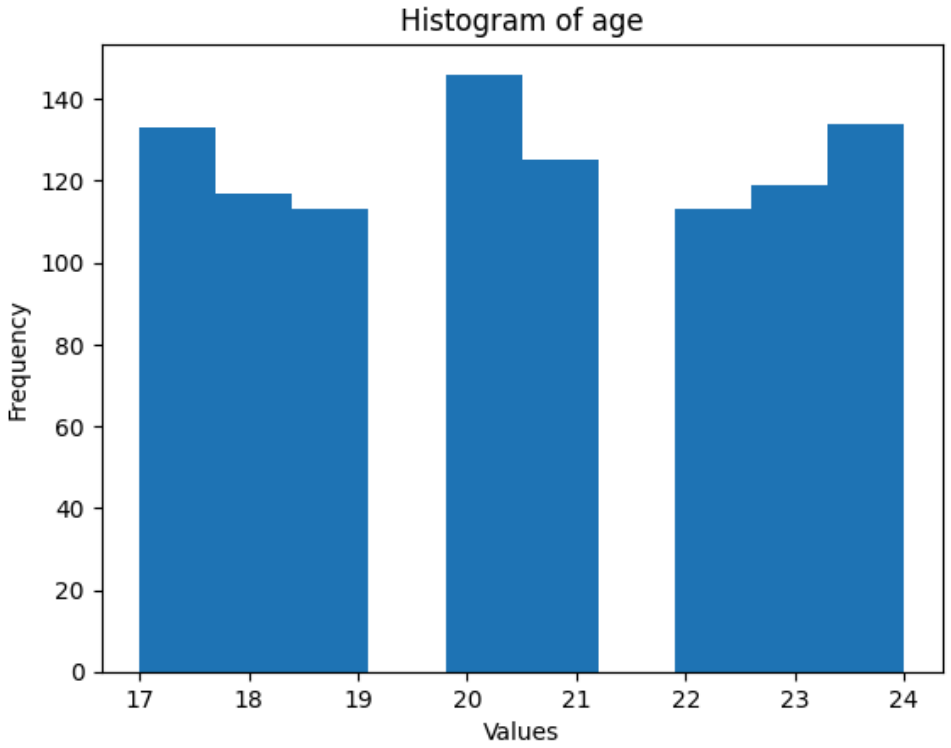

Boxplot of mental_health_rating
Values
2
4
6
8
10
1

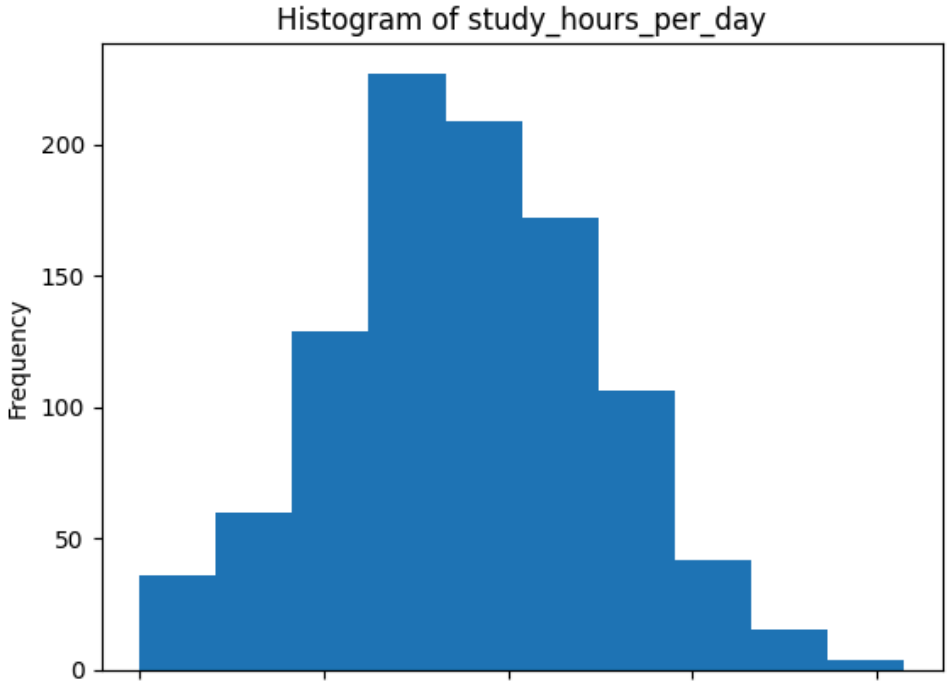

Boxplot of exam_score
Values
100
90
80
70
60
50
40
30
20
1

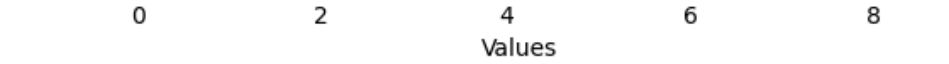
Haz doble clic (o pulsa Intro) para editar

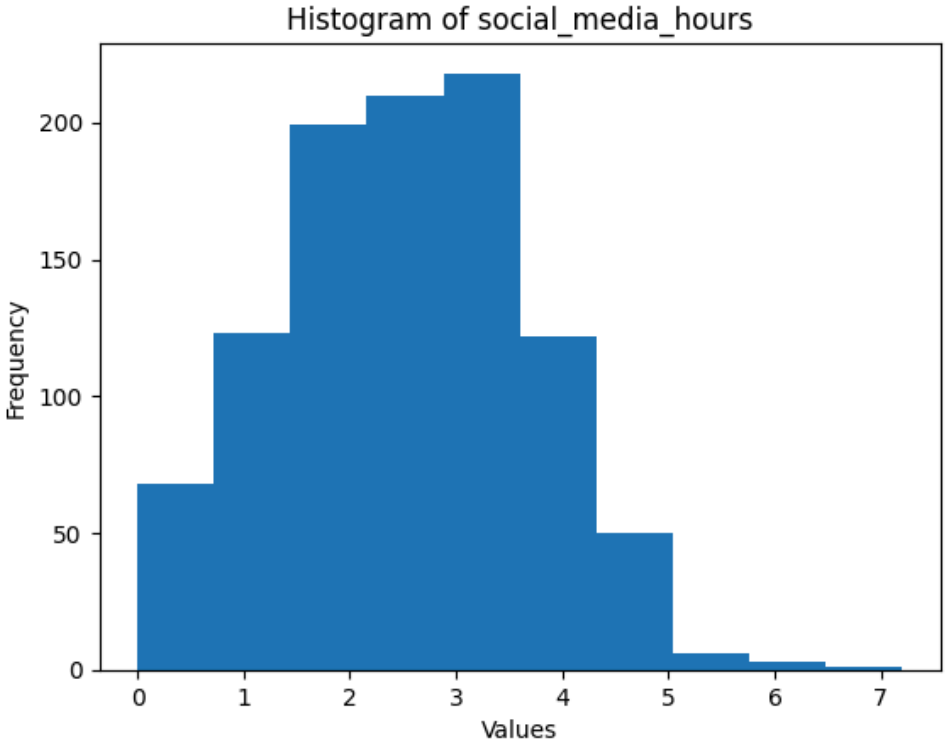
Histograms

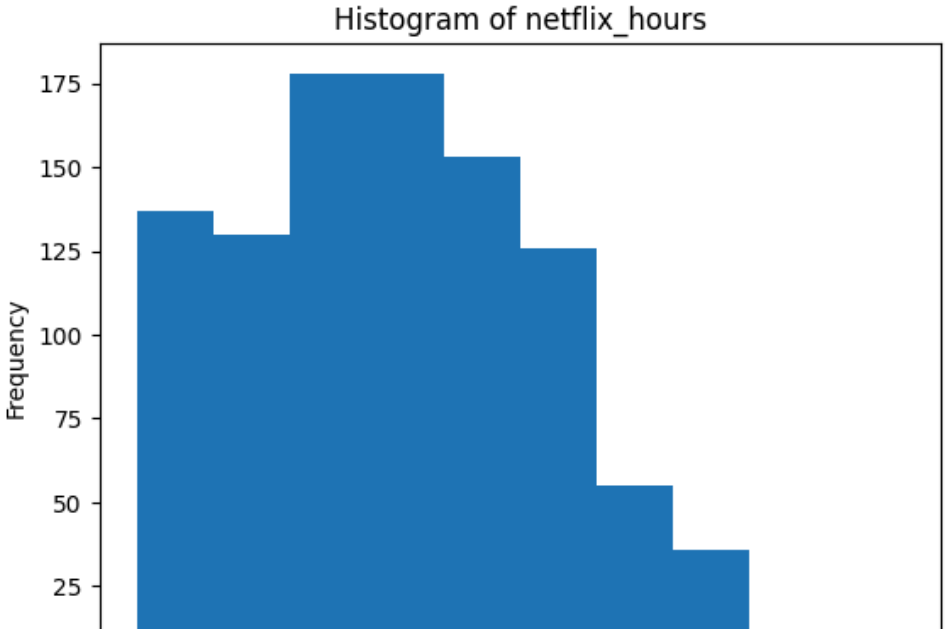
```
for col in col_names:
  plt.hist(numeric_data[col])
  plt.title(f'Histogram of {col}')
  plt.xlabel('Values')
  plt.ylabel('Frequency')
  plt.show()
```

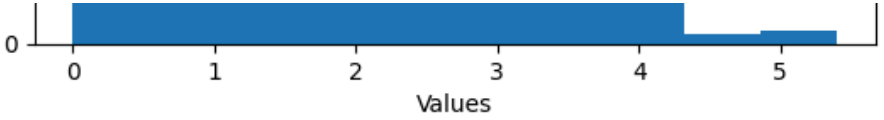

Histogram of age
Frequency
Values
0
20
40
60
80
100
120
140
17
18
19
20
21
22
23
24

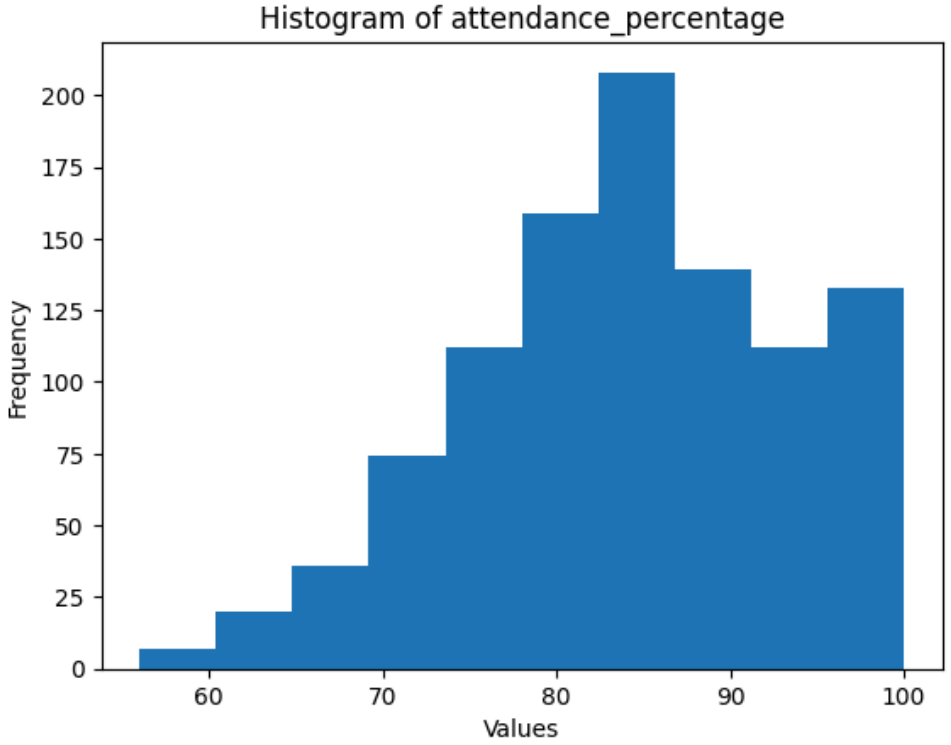

Histogram of study_hours_per_day
Frequency
0
50
100
150
200

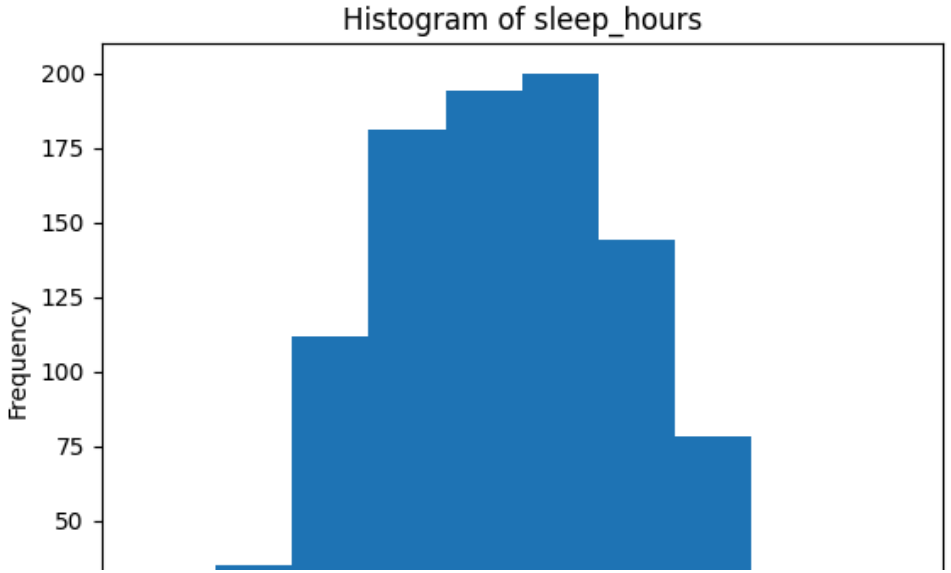

0
2
4
6
8
Values

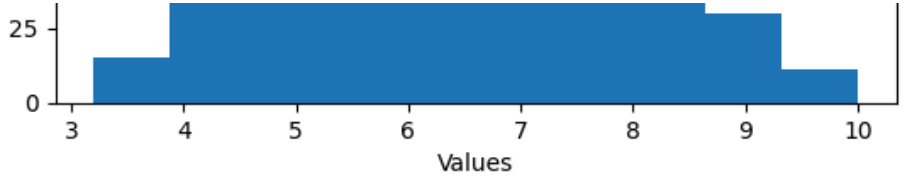

Histogram of social_media_hours
Frequency
0
50
100
150
200
0
1
2
3
4
5
6
7
Values

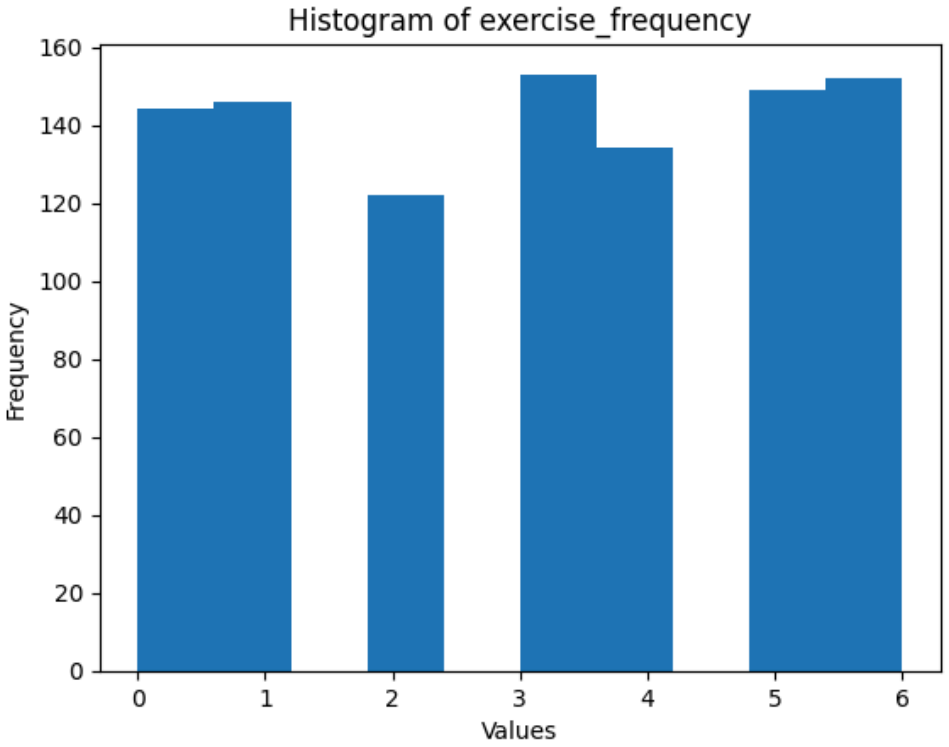

Histogram of netflix_hours
Frequency
25
50
75
100
125
150
175

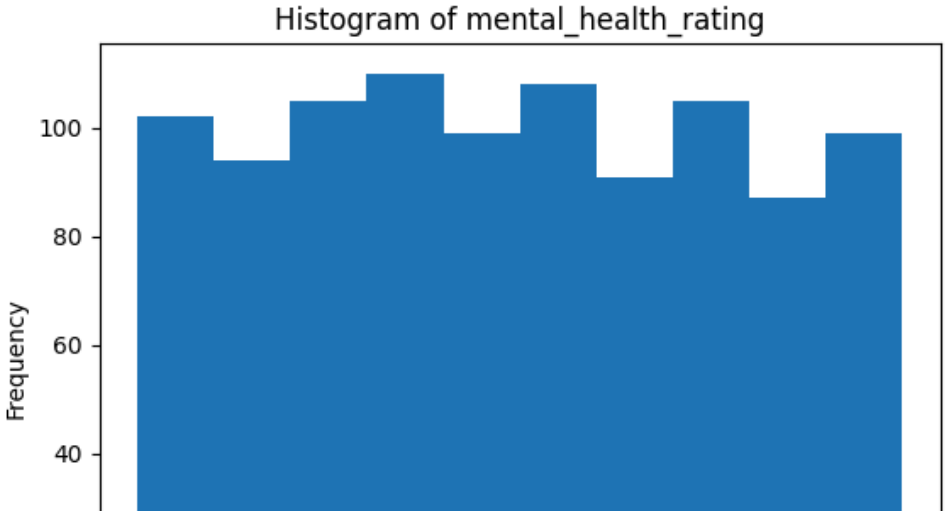

0
0
1
2
3
4
5
Values

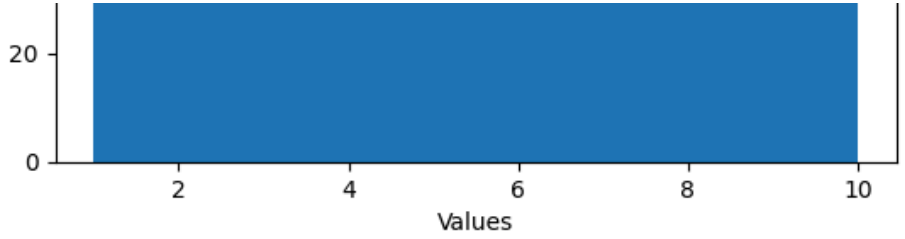

Histogram of attendance_percentage
Frequency
200
175
150
125
100
75
50
25
0
60
70
80
90
100
Values

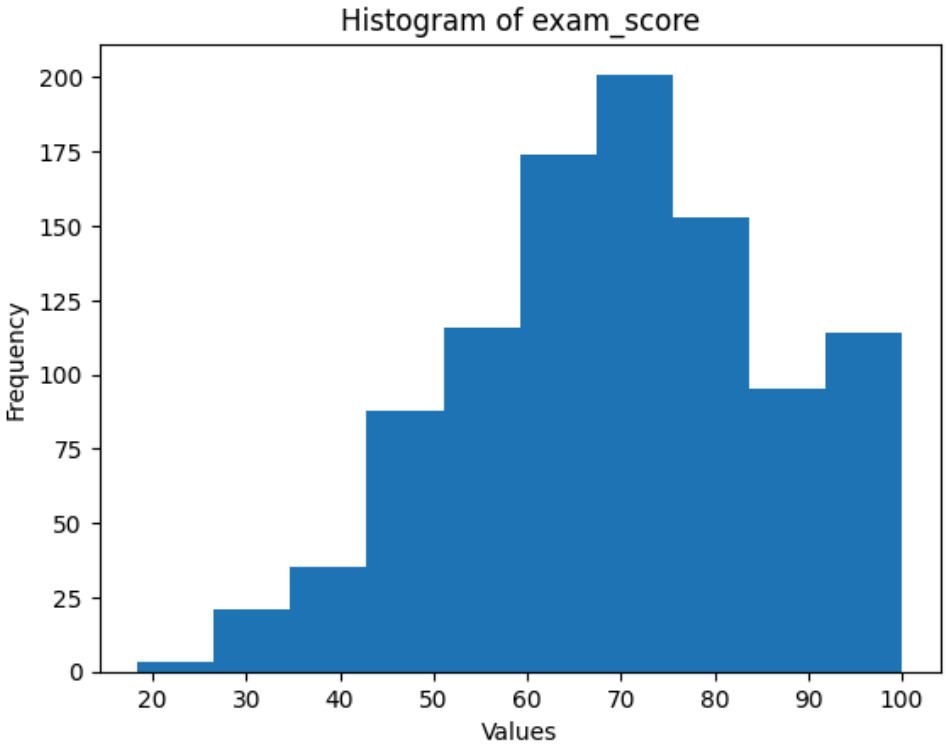

Histogram of sleep_hours
Frequency
200
175
150
125
100
75
50


25
0
3
4
5
6
7
8
9
10
Values


Histogram of exercise_frequency
160
140
120
100
80
60
40
20
0
Frequency
0
1
2
3
4
5
6
Values


Histogram of mental_health_rating
100
80
60
40
Frequency


20
0
2
4
6
8
10
Values


Histogram of exam_score
Frequency
200
175
150
125
100
75
50
25
0
20
30
40
50
60
70
80
90
100
Values

Función de correlación

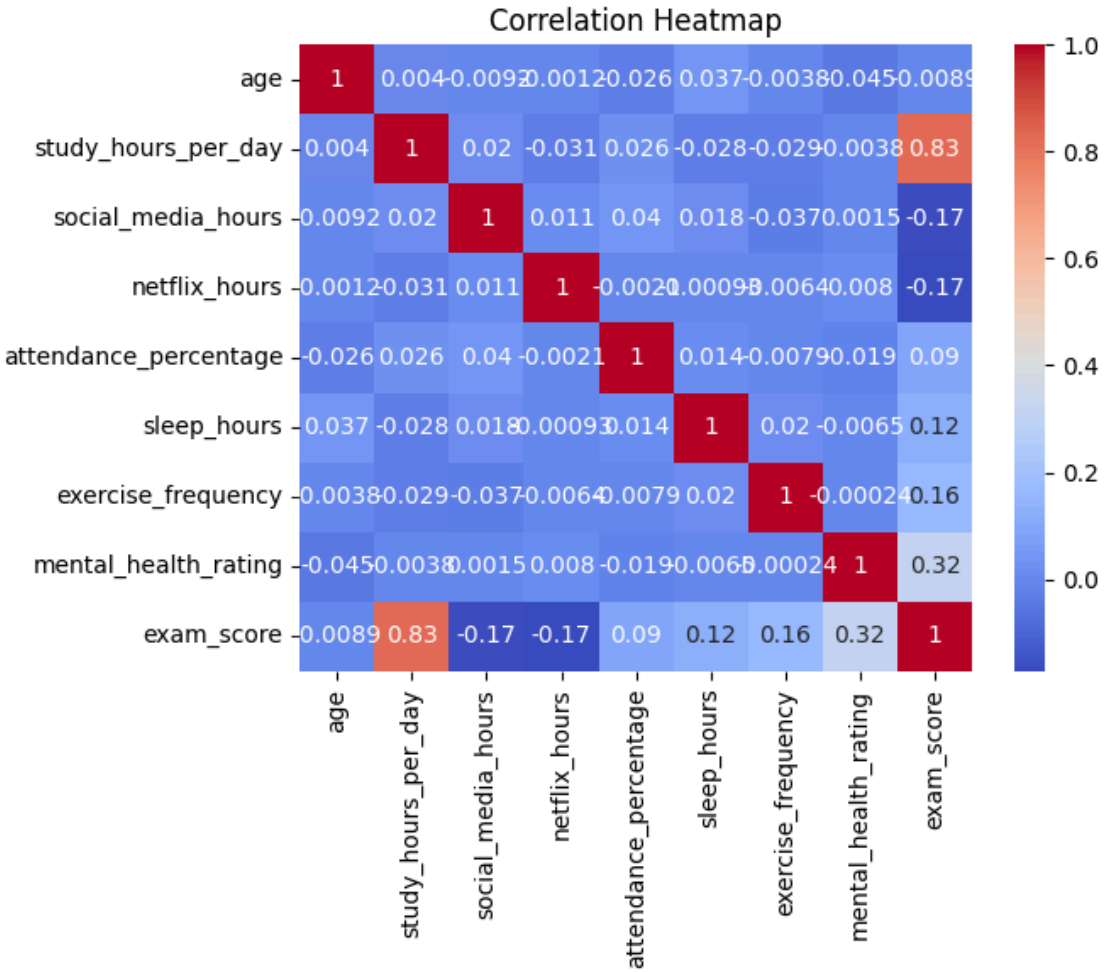
```
num_data_corr = numeric_data.corr()
print(num_data_corr)
```

```
                            age  study_hours_per_day  social_media_hours  \
age                    1.000000             0.003971           -0.009151
study_hours_per_day    0.003971             1.000000            0.020282
social_media_hours    -0.009151             0.020282            1.000000
netflix_hours         -0.001174            -0.031158            0.011477
attendance_percentage -0.026055             0.026264            0.040479
sleep_hours            0.037482            -0.027757            0.018236
exercise_frequency    -0.003836            -0.028701           -0.037319
mental_health_rating  -0.045101            -0.003768            0.001496
exam_score            -0.008907             0.825419           -0.166733

                       netflix_hours  attendance_percentage  sleep_hours  \
age                        -0.001174              -0.026055     0.037482
study_hours_per_day        -0.031158               0.026264    -0.027757
social_media_hours          0.011477               0.040479     0.018236
netflix_hours               1.000000              -0.002092    -0.000935
attendance_percentage      -0.002092               1.000000     0.013756
sleep_hours                -0.000935               0.013756     1.000000
exercise_frequency         -0.006448              -0.007857     0.019769
mental_health_rating        0.008034              -0.018745    -0.006508
exam_score                 -0.171779               0.089836     0.121683

                       exercise_frequency  mental_health_rating  exam_score
age                             -0.003836             -0.045101   -0.008907
study_hours_per_day             -0.028701             -0.003768    0.825419
social_media_hours              -0.037319              0.001496   -0.166733
netflix_hours                   -0.006448              0.008034   -0.171779
attendance_percentage           -0.007857             -0.018745    0.089836
sleep_hours                      0.019769             -0.006508    0.121683
exercise_frequency               1.000000             -0.000242    0.160107
mental_health_rating            -0.000242              1.000000    0.321523
exam_score                       0.160107              0.321523    1.000000
```

Heatmap

```
sns.heatmap(num_data_corr, annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
```

```
plt.show()
```


Correlation Heatmap
age
study_hours_per_day
social_media_hours
netflix_hours
attendance_percentage
sleep_hours
exercise_frequency
mental_health_rating
exam_score
1.0
0.8
0.6
0.4
0.2
0.0