Tecnológico de Monterrey

Campus Ciudad de México

Activity 3: Patterns with K-means

Ximena Silva Bárcena

A01785518

Rodrigo Martínez Vallejo

A00573055

Prof. Sergio Ruiz Loza

Mastering Analytics

Grupo **201**

Wednesday, May 7th, 2025

Link of google colab:

https://colab.research.google.com/drive/1U31gvCgORY4tQIheh5GA6qPi3iuKo4fu?usp=sharing

Ignored Variables:

Student_id: is a unique identifier and does not contribute to understanding the relationships between study habits and academic success.

Gender: is a categorical variable.

Sleep-hour: show low correlation with the exam_score and other key performance indicators.

Diet_quality: is a categorical variable.

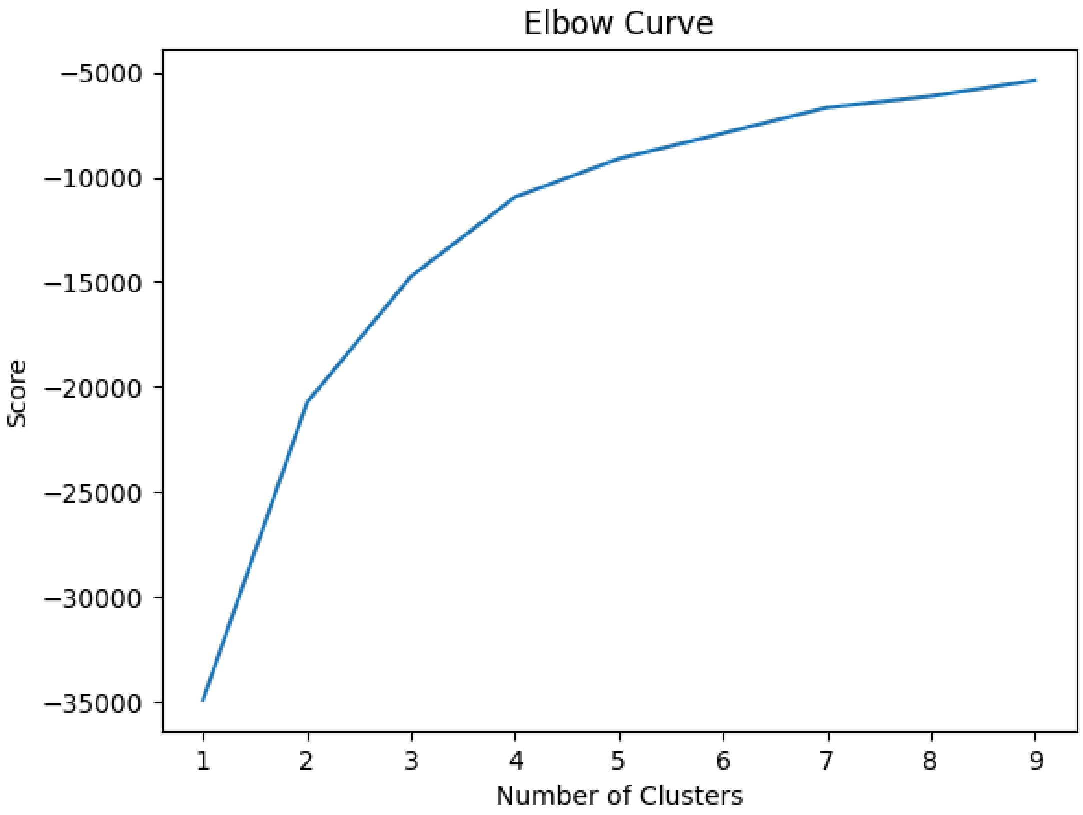
Extracurricular_participation: low correlation with the exam_score.

Internet_quality: low correlation with exam_score.

Parental_education_level: low correlation with exam_score.

Value for k: 3

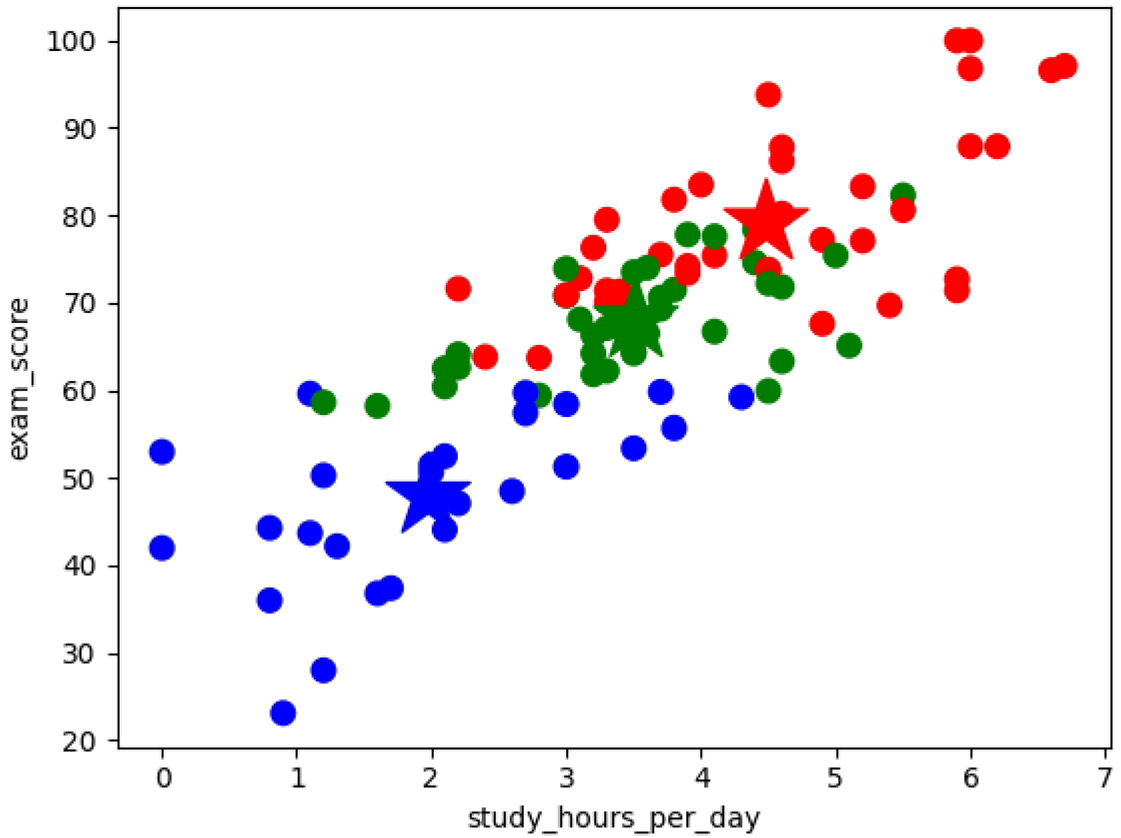
The value of K was deceived upon the review of the Elbow Curve, from which it was decided to stick with a value of 3.



Centroids of K-means:

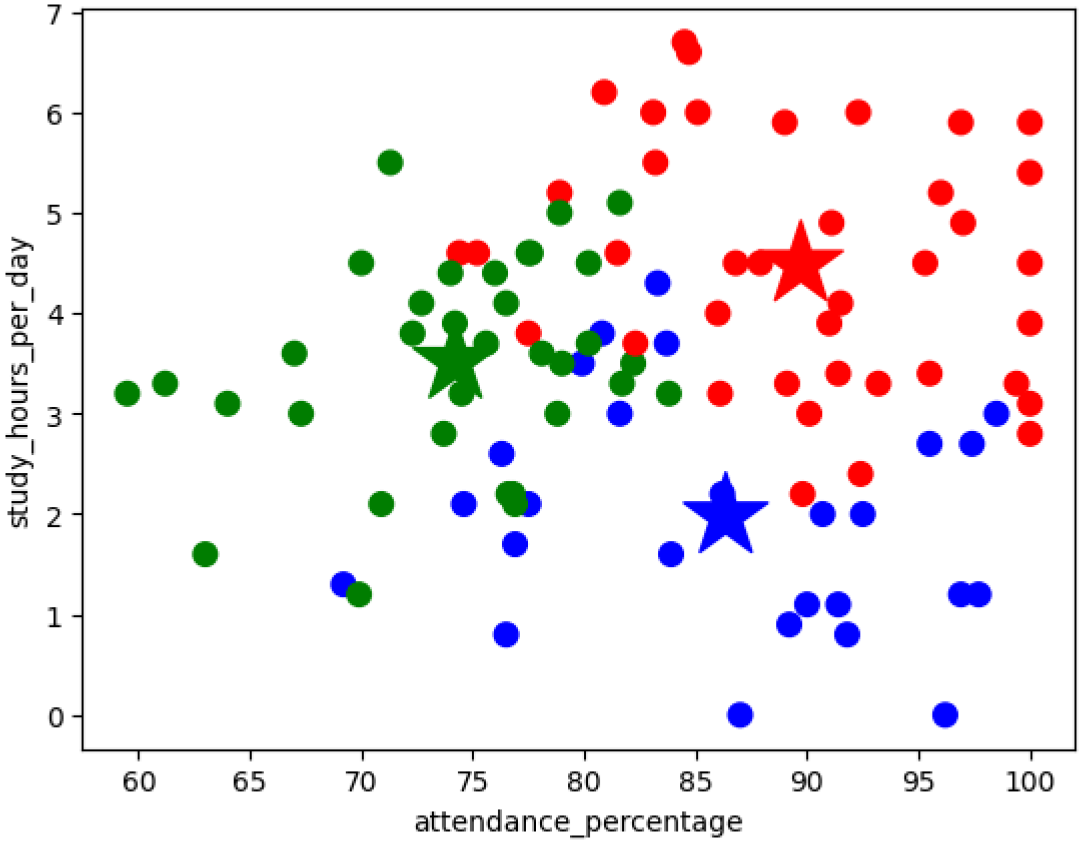
- Exam scores vs Study hours per day

This scatter plot reveals a strong positive correlation between the number of hours students dedicate to studying and their exam performance. The red cluster is concentrated around 4.5 study hours per day and plus 80 in the exam scores, suggesting that increased study time pays off significantly. The green cluster falls around 3.5 hours of study and 65 as result, while the blue cluster shows the least study time at 2 hours of study and scores around 45 and 50.



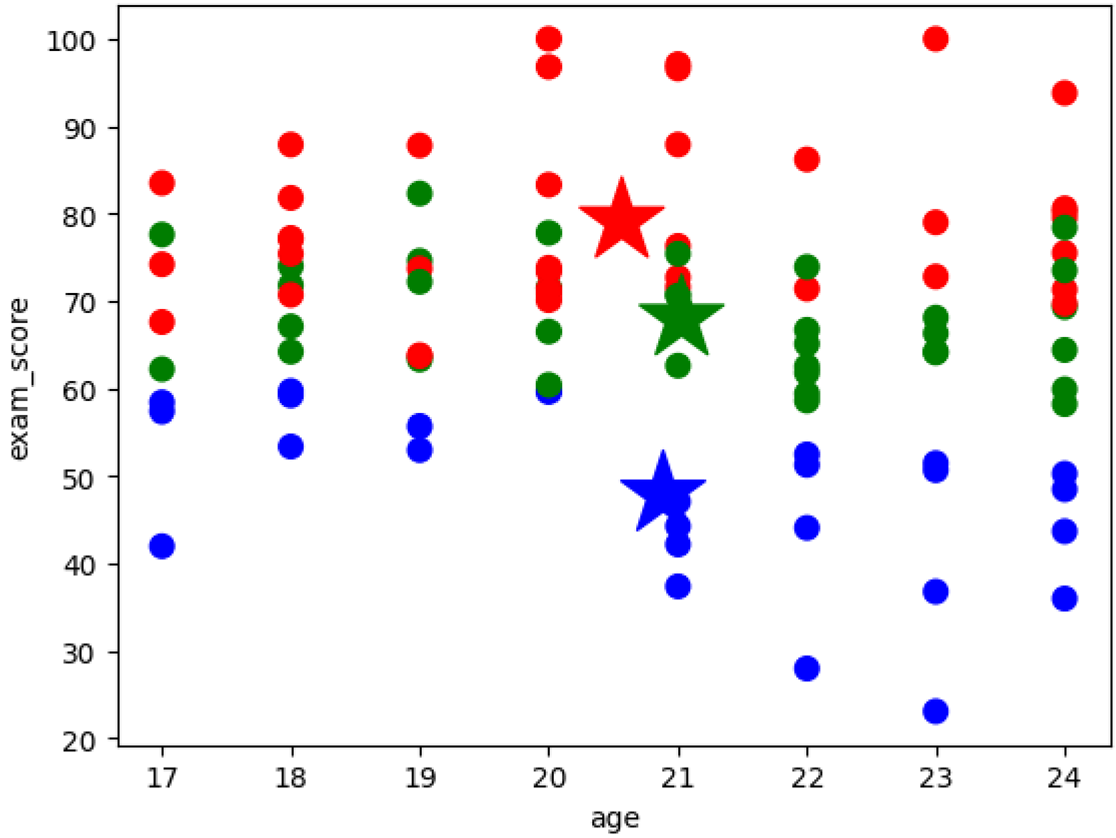
- Study hours per day vs Attendance percentage

The clusters of this second graph correlate perfectly to the information obtained from the first one, where the red cluster corresponds to both the ones who studied the most and the ones who attended the majority of the classes. This revelation reinforces the idea that discipline corresponds to higher scores. On the other hand, the green cluster balances moderate attendance with moderate study time, and the blue cluster, despite decent attendance, studies far less, showing that attendance alone isn't enough without active study time.



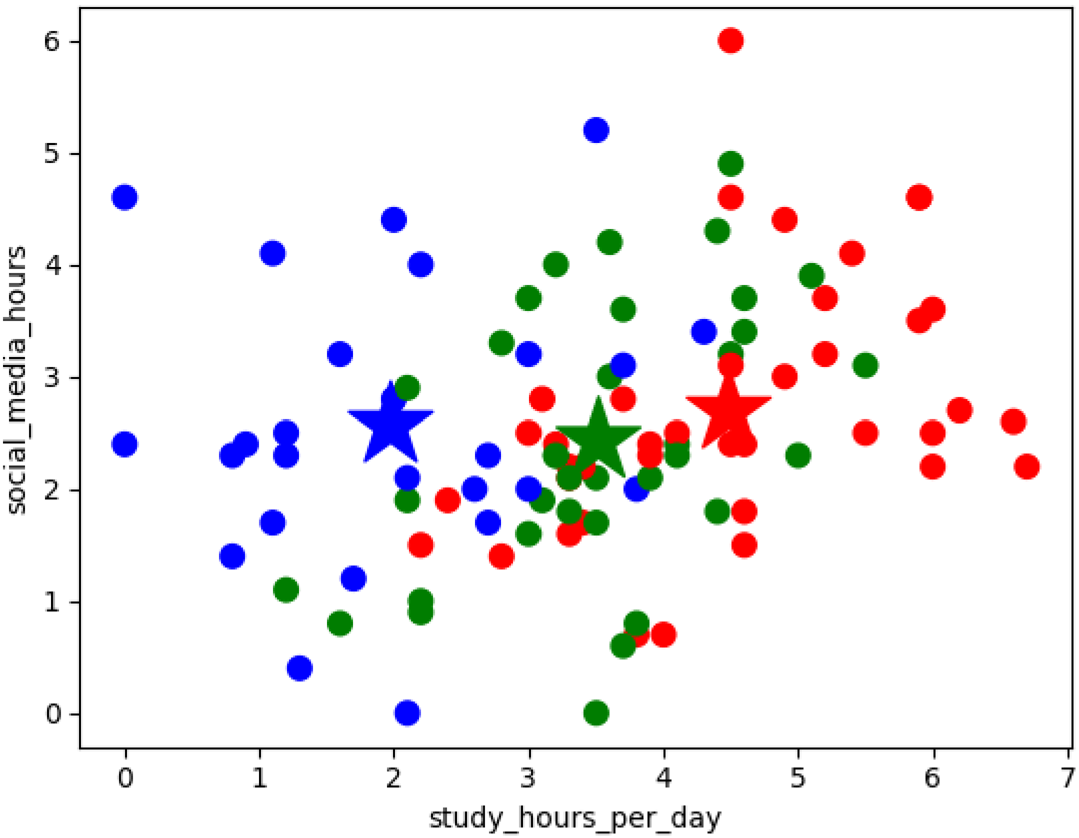
- Exam scores vs Age

The centroids in this graph reveal that age is not a factor to take into account regarding academic performance, as the age gap is distributed pretty much equally across the three groups.



- Social media hours vs Study hours per day

In this chart it can be observed that the blue cluster spends a similar amount of time on social media despite studying less. The red cluster balances study and screen time more effectively, clustering around 4.5 study hours of study with pretty much the same social media time as the other two clusters. From this it could be said that social media hours of use don't appear to play a significant role in this case regarding the ability to spend the same time studying or more.



Clusters:

- Cluster 1: These students show high academic performance with consistently high study hours, high attendance, and moderate screen time. They likely have strong time management and learning strategies.
- Cluster 2: This group maintains average study time, moderate attendance, and average exam performance. They may benefit from improved study habits or targeted academic support.
- Cluster 3: Students in this cluster attend class fairly regularly but study significantly less and perform poorly. Their screen time is pretty much the same as the other clusters.

Questions:

Ximena Silva Bárcena
Do you think these centers might be representative of the data? Why?

Yes, these centers seem representative because they capture the main patterns in the data. Each cluster center reflects a unique combination of student habits and performance. For instance, one cluster has higher study hours and exam scores, while another has lower study hours and mental health rating.
How did you obtain the k value to use?
With the Elbow Method, which involves plotting the within-cluster sum of squares (WCSS) for different values of k. We chose k=3 because the curve showed a noticeable bend at this point, indicating a significant reduction in WCSS
Would the centers be more representative if you used a higher value? A lower value?
Using a higher k would produce more granular clusters, potentially capturing more nuanced differences in student behaviors, but it risks creating overly specific groups that might not generalize well. Conversely, a lower k would result in broader, less precise groupings.
How far apart are the centers? Are any very close to others?
The centers are relatively distinct, as seen in their age, study hours, and exam score differences. However, two clusters (Cluster 1 and Cluster 2) have relatively similar attendance_percentage and sleep_hours values, indicating that these groups may share some common habits despite other differences.
What would happen to the centers if we had many outliers in the box-and-whisker analysis?
Outliers can significantly distort the cluster centers by pulling them toward extreme values, leading to less accurate and less representative clusters. This impact is especially problematic if the outliers represent rare or exceptional cases that do not reflect typical student behavior
What can you say about the data based on the centers?

Cluster 0: Lower study hours, lower mental health ratings, moderate exercise, and the lowest exam scores, possibly indicating students struggling academically.

Cluster 1: Moderate study hours, average mental health, and higher exam scores, potentially representing balanced students.
Cluster 2: Higher study hours, better exercise habits, and the highest exam scores, likely indicating high-performing students with a strong focus on academics.

Rodrigo Martínez Vallejos
Do you think these centers might be representative of the data? Why?
Centers are likely representative because they capture distinct groupings in the data, reflecting different student profiles. For example, one cluster has higher study hours and exam scores, suggesting a more academically focused group, while another has lower study hours and moderate exam scores, potentially indicating a more balanced lifestyle.
How did you obtain the k value to use?
We used the elbow method which is based on the principle that adding more clusters reduces the total within-cluster variance but with diminishing returns after a certain point. The "elbow" is this point of diminishing returns in our case k = 3.
Would the centers be more representative if you used a higher value? A lower value?
Higher k might overfit the data, capturing noise rather than true patterns, while a lower k would likely underfit, grouping distinct student behaviors into overly broad

clusters. This would obscure meaningful differences, like those seen in exercise frequency and mental health ratings.
How far apart are the centers? Are any very close to others?
The centers show differences in attributes like study hours, exercise frequency, and exam scores. However, the age and attendance percentage are relatively similar, suggesting that these features may have less impact on the cluster differentiation.
What would happen to the centers if we had many outliers in the box-and-whisker analysis?
Outliers distort the cluster centers, pulling them away from the true central tendencies of their respective groups. This would lead to less accurate representations of the average student in each cluster, potentially misleading subsequent analysis.
What can you say about the data based on the centers?
The data appears to contain three main student profiles: one with balanced study habits and average scores, one focused on high academic performance with more study hours and exercise, and another with lower academic focus and moderate exam performance

Patterns with K means

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

```
df = pd.read_csv('student_habits_performance.csv')
```

```
df.info()
```

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

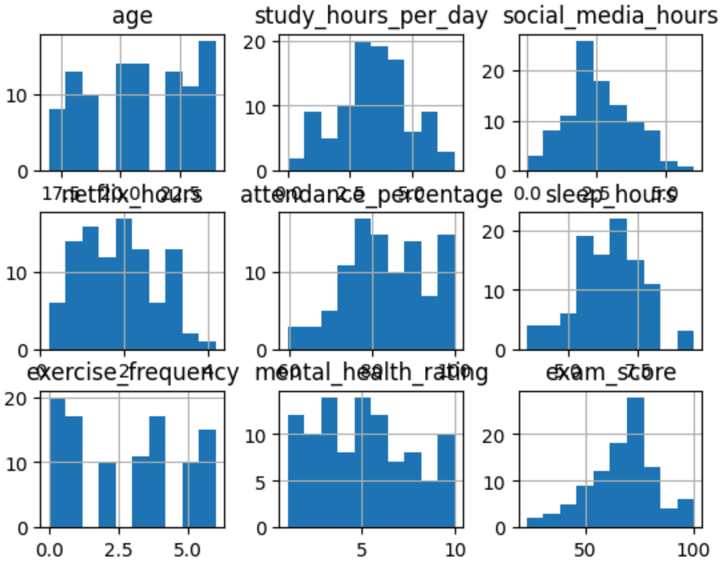
```
dtypes: float64(6), int64(3), object(7)
memory usage: 125.1+ KB
```

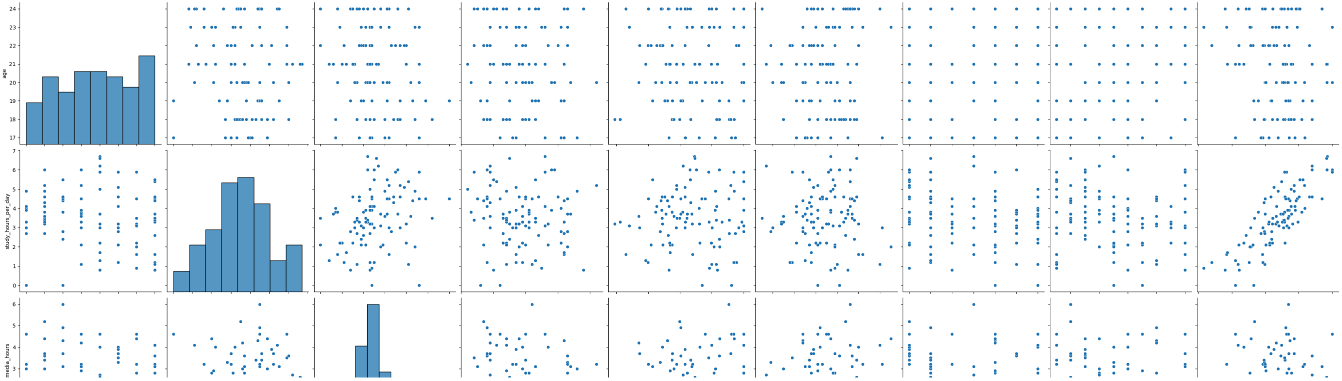
```
selected_columns = ["age", "study_hours_per_day", "social_media_hours", "netflix_hours", "attendance_percentage", "sleep_hours", "exercis
```

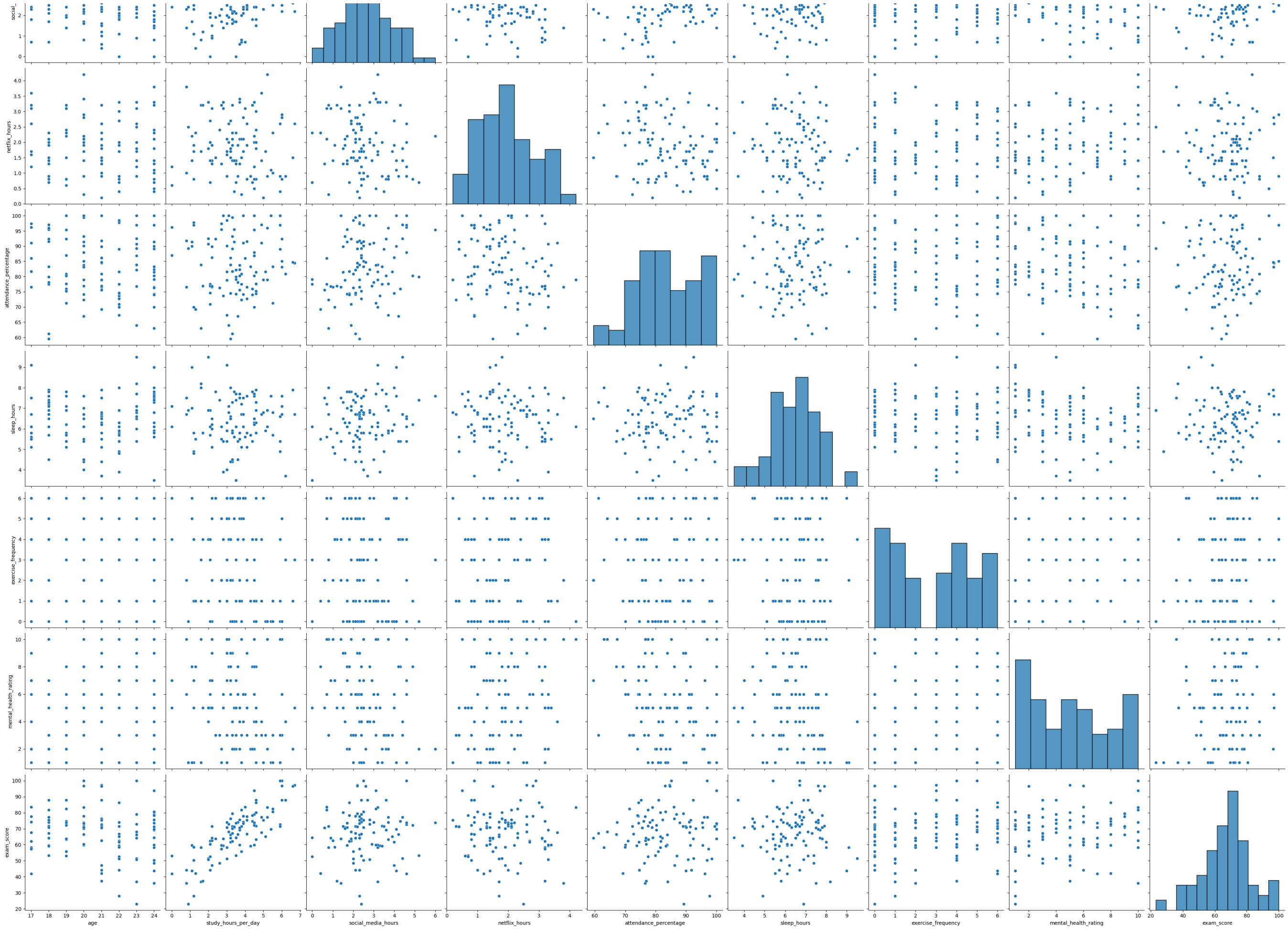
```
#threshold = df['study_hours_per_day'].quantile(0.75)
#filtered_df = df[df['study_hours_per_day'] >= threshold]
#dataframe_selected = filtered_df[selected_columns]

filtered_df = df.sample(frac=0.1, random_state=42)
dataframe_selected = filtered_df[selected_columns]
```

Histogram

```
dataframe_selected.hist()
plt.show()

# Pairplot
sb.pairplot(dataframe_selected, height=4, kind='scatter')
plt.show()
```


age
study_hours_per_day
social_media_hours
netflix_hours
attendance_percentage
sleep_hours
exercise_frequency
mental_health_rating
exam_score
0
10
20
5
17.5
20.0
22.5
0.0
2.5
5.0
0
2
4
60
80
100
7.5
10
50


age
study_hours_per_day
media_hours
17
18
19
20
21
22
23
24
0
1
2
3
4
5
6
7

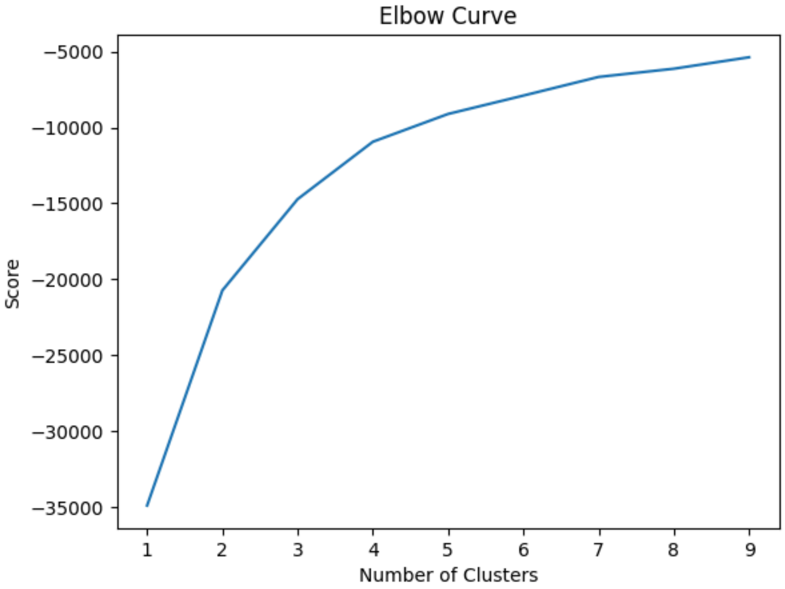


social_
0
1
2
netflix_hours
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
attendance_percentage
60
65
70
75
80
85
90
95
100
sleep_hours
4
5
6
7
8
9
exercise_frequency
0
1
2
3
4
5
6
mental_health_rating
2
4
6
8
10
exam_score
20
30
40
50
60
70
80
90
100
17 18 19 20 21 22 23 24
age
0 1 2 3 4 5 6 7
study_hours_per_day
0 1 2 3 4 5 6
social_media_hours
0 1 2 3 4
netflix_hours
60 70 80 90 100
attendance_percentage
4 5 6 7 8 9
sleep_hours
0 1 2 3 4 5 6
exercise_frequency
2 4 6 8 10
mental_health_rating
20 40 60 80 100
exam_score

K-Means with chosen K value

```
# Matriz de entrada
X = np.array(dataframe_selected)
print("Shape:", X.shape)

# Elbow curve
Nc = range(1, 10)
kmeans_models = [KMeans(n_clusters=i, n_init=10) for i in Nc]
scores = [model.fit(X).score(X) for model in kmeans_models]
plt.plot(Nc, scores)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

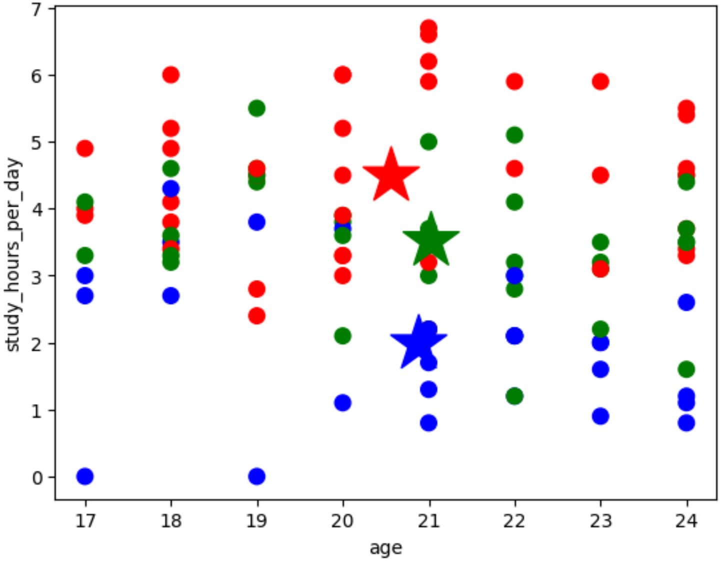
```
Shape: (100, 9)
```



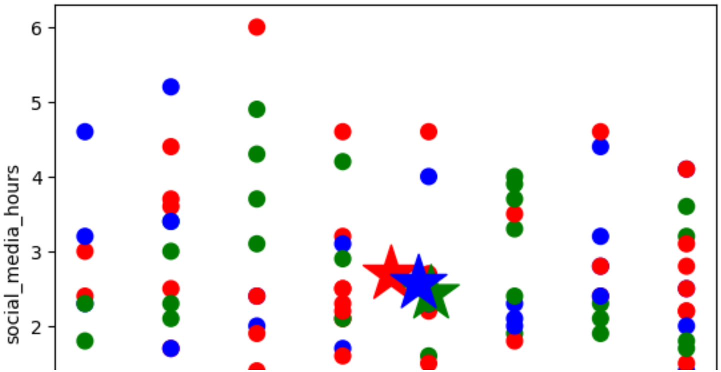
Haz doble clic (o pulsa Intro) para editar

```
# Elegir K = 4
kmeans = KMeans(n_clusters=3, n_init=10).fit(dataframe_selected)
labels = kmeans.predict(dataframe_selected)
centroids = kmeans.cluster_centers_

# Asignar colores
colores = ['red','green', 'blue']
asignar = [colores[i] for i in labels]
```

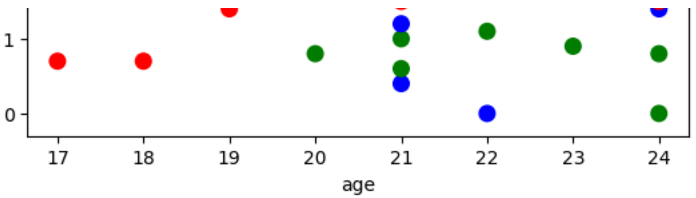
```
for i in range(len(selected_columns)):
    for j in range(len(selected_columns)):
        if i != j:
            plt.scatter(X[:, i], X[:, j], c=asignar, s=70)
            plt.scatter(centroids[:, i], centroids[:, j], marker="*", c=colores, s=1000)
            plt.xlabel(selected_columns[i])
            plt.ylabel(selected_columns[j])
            plt.show()
```

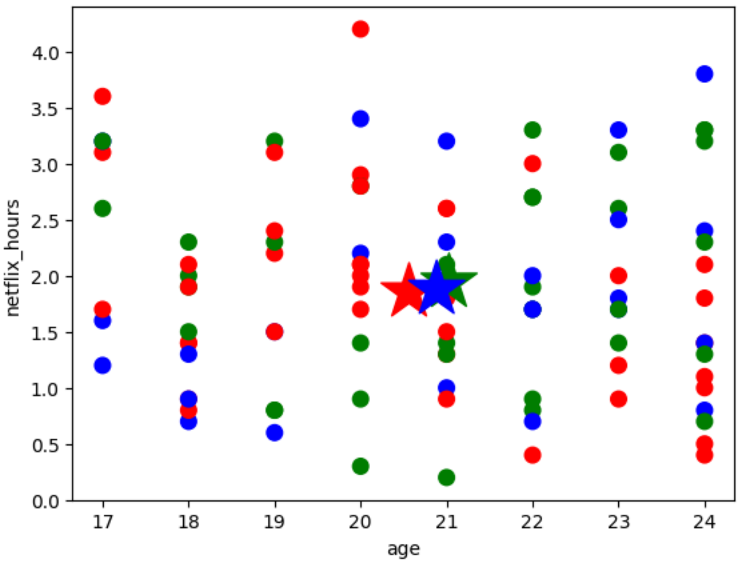


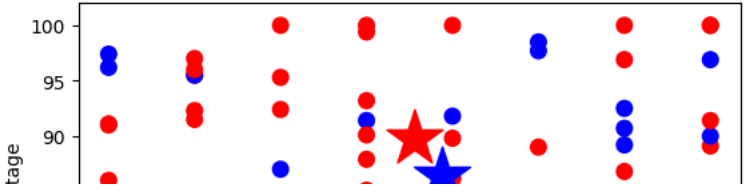
study_hours_per_day
7
6
5
4
3
2
1
0
17
18
19
20
21
22
23
24
age

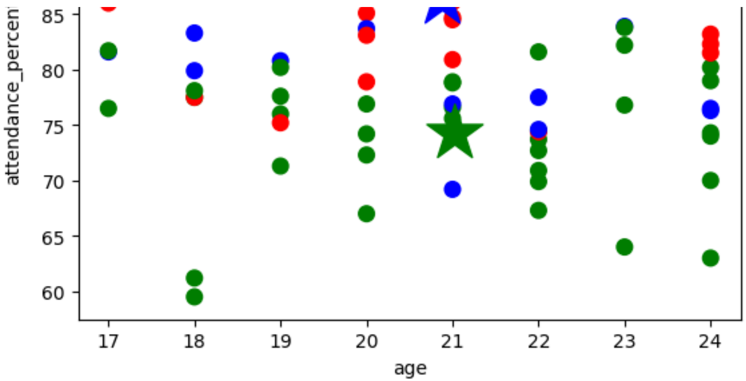


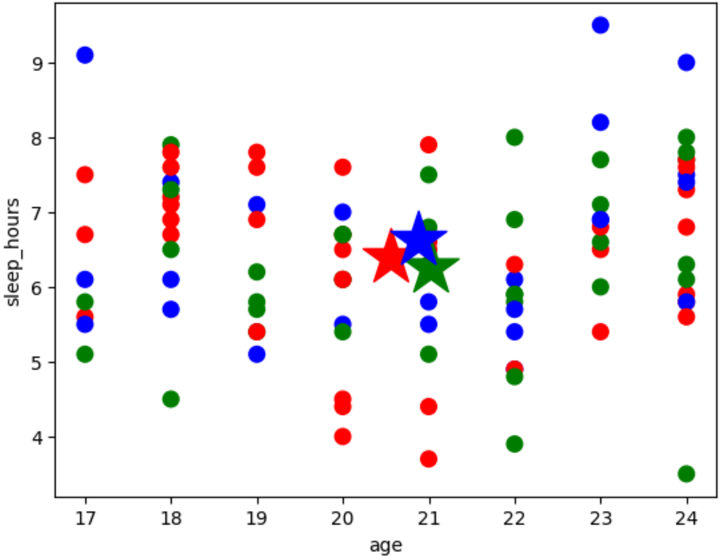
social_media_hours
6
5
4
3
2

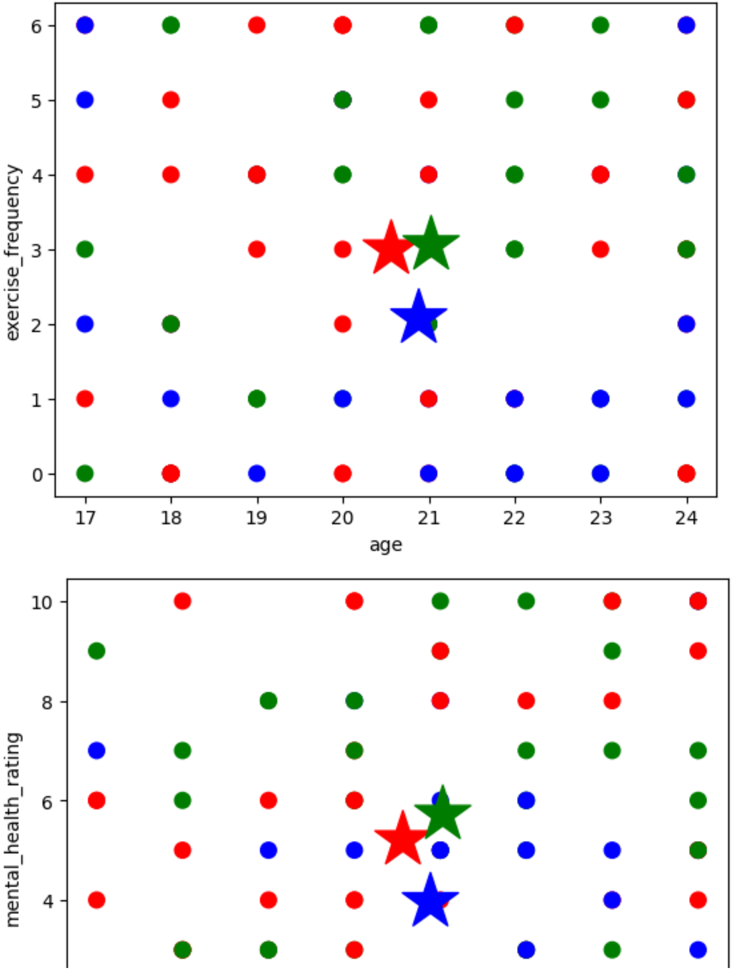

1
0
17
18
19
20
21
22
23
24
age

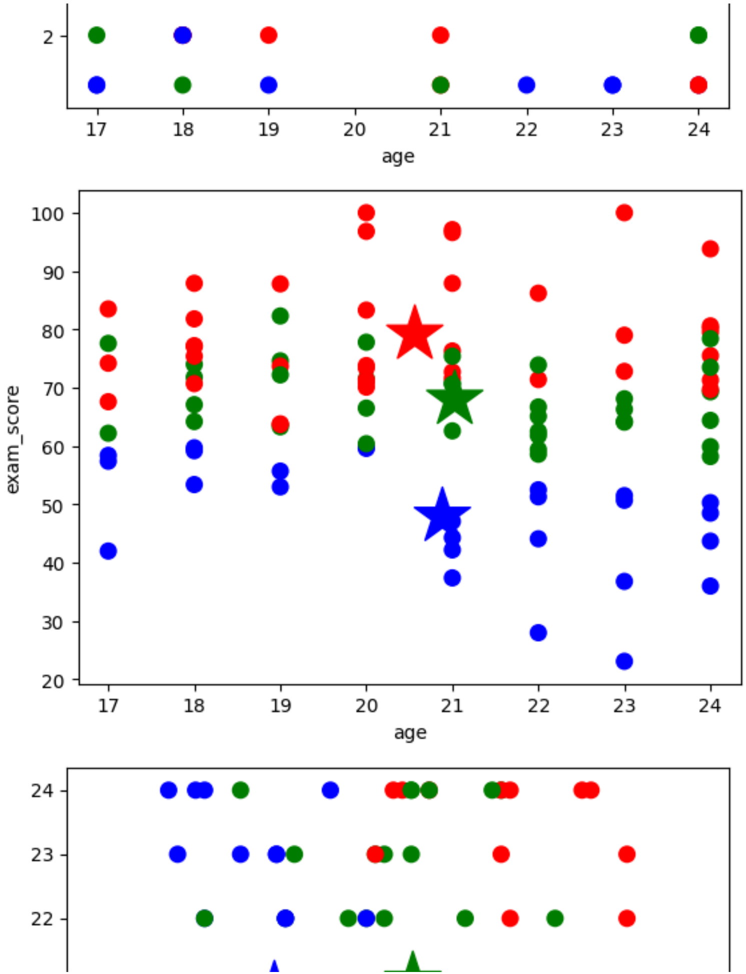

4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
netflix_hours
17
18
19
20
21
22
23
24
age

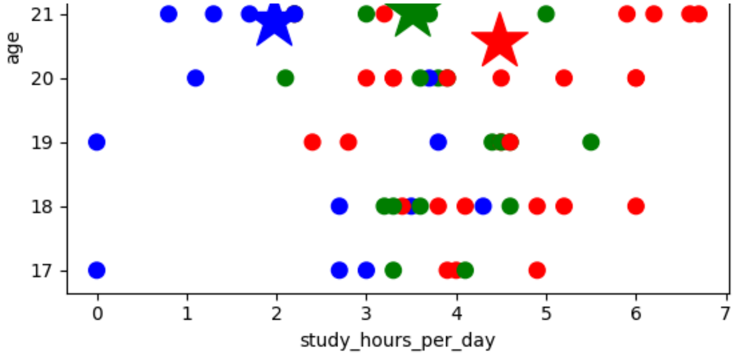

100
95
90
tage

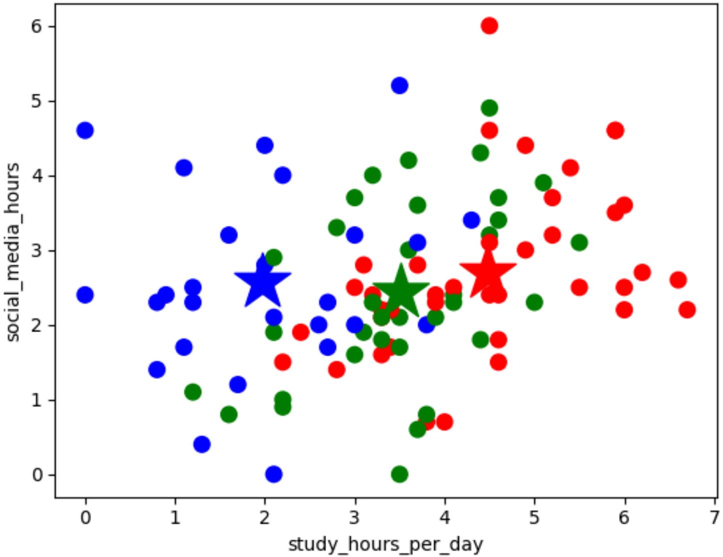

attendance_percen
85
80
75
70
65
60
17
18
19
20
21
22
23
24
age

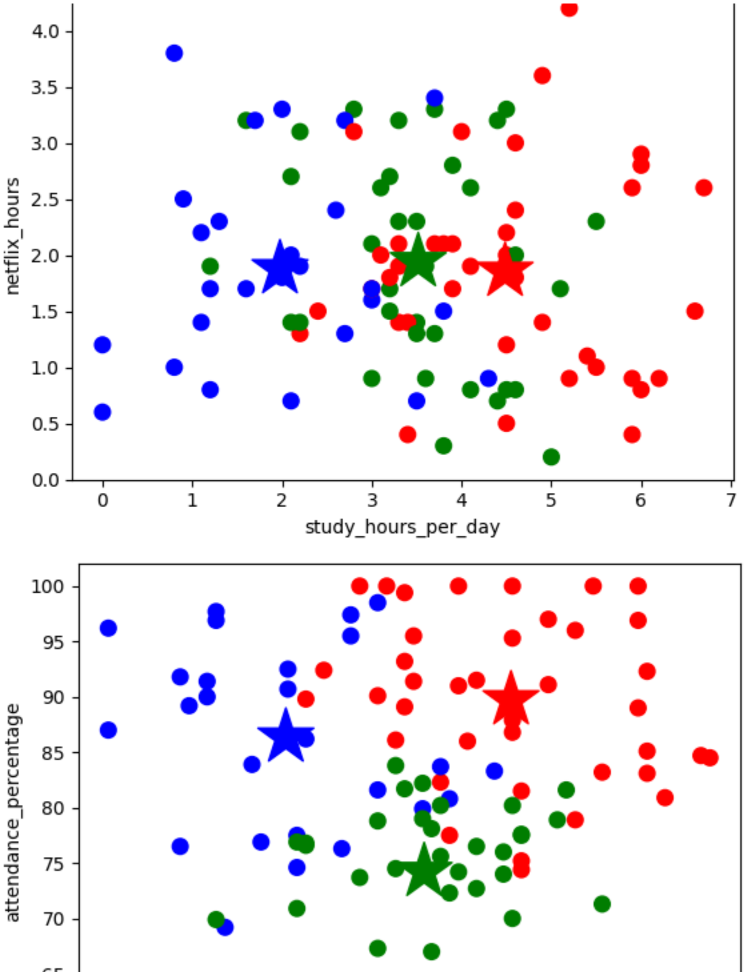

sleep_hours
9
8
7
6
5
4
17
18
19
20
21
22
23
24
age


exercise_frequency
0
1
2
3
4
5
6
17
18
19
20
21
22
23
24
age
mental_health_rating
4
6
8
10

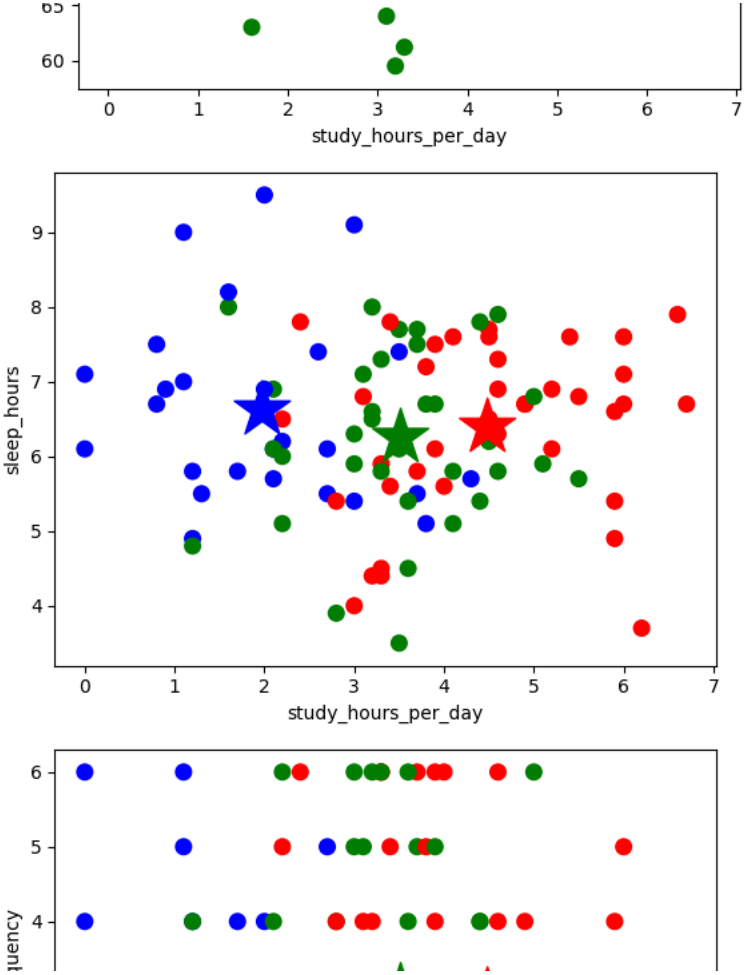

2
17
18
19
20
21
22
23
24
age
100
90
80
70
60
50
40
30
20
exam_score
17
18
19
20
21
22
23
24
age
24
23
22

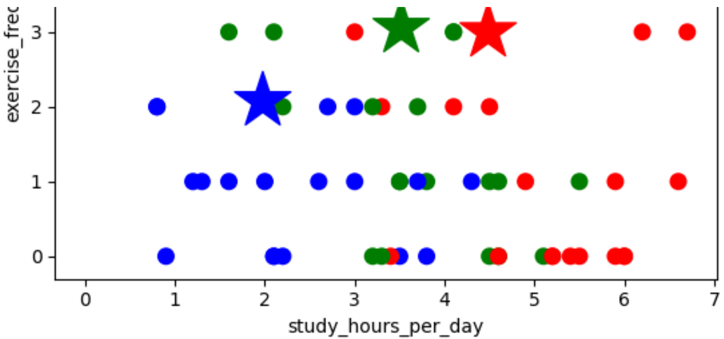

age
21
20
19
18
17
0
1
2
3
4
5
6
7
study_hours_per_day

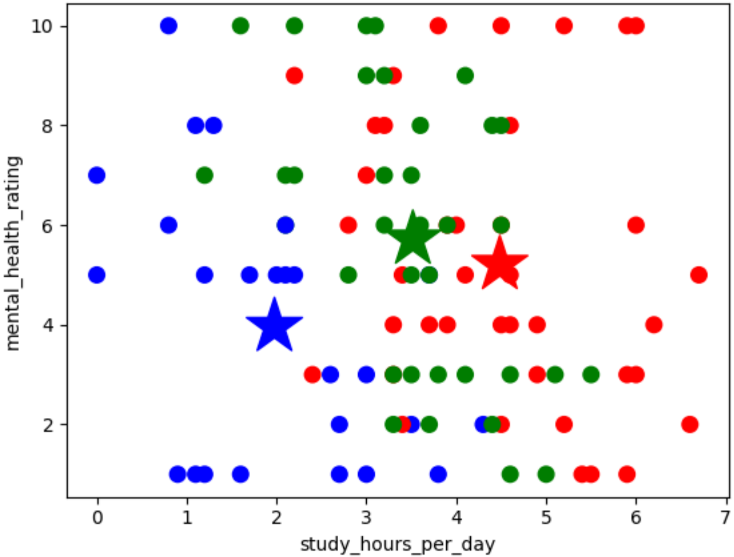

social_media_hours
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
study_hours_per_day



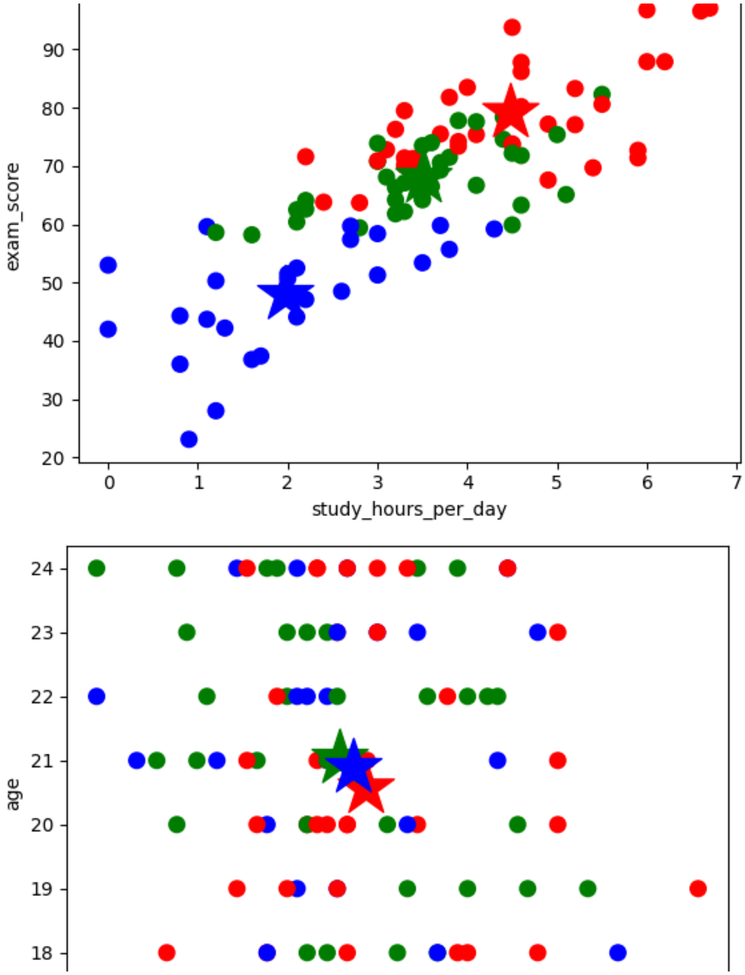
netflix_hours
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
study_hours_per_day
attendance_percentage
100
95
90
85
80
75
70
65

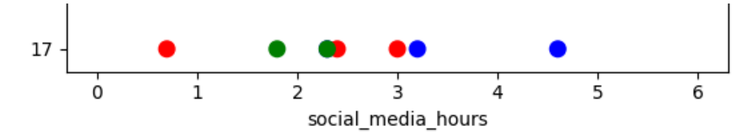

65
60
0
1
2
3
4
5
6
7
study_hours_per_day
sleep_hours
9
8
7
6
5
4
0
1
2
3
4
5
6
7
study_hours_per_day
6
5
4
quency

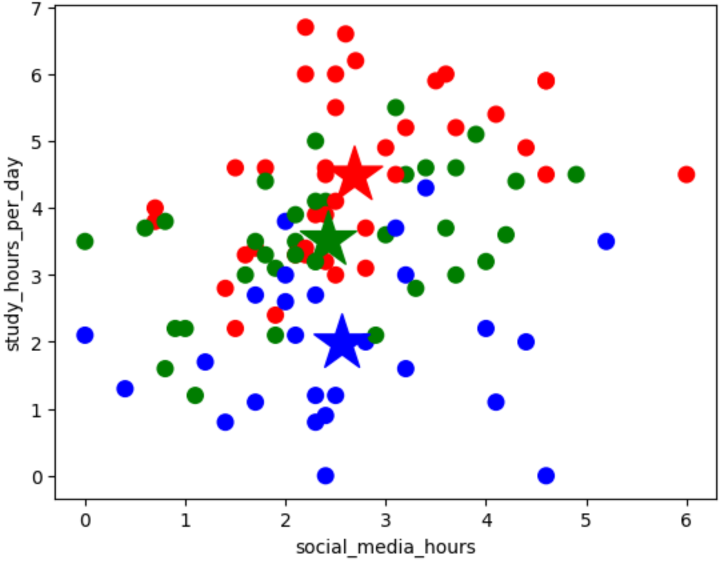

exercise_frec
0
1
2
3
0
1
2
3
4
5
6
7
study_hours_per_day

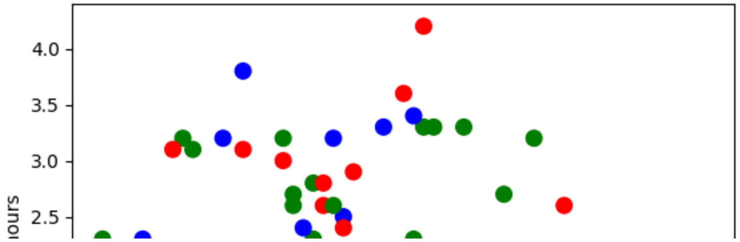

mental_health_rating
2
4
6
8
10
0
1
2
3
4
5
6
7
study_hours_per_day

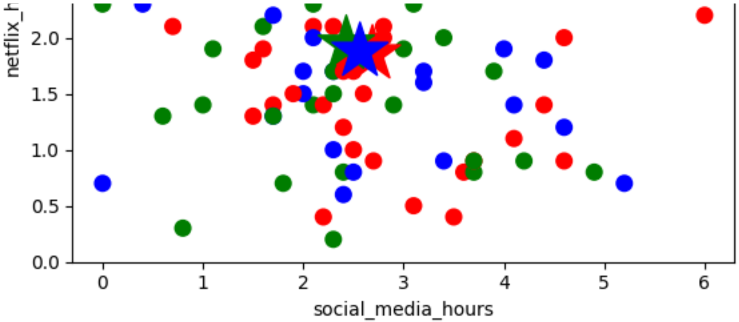
100

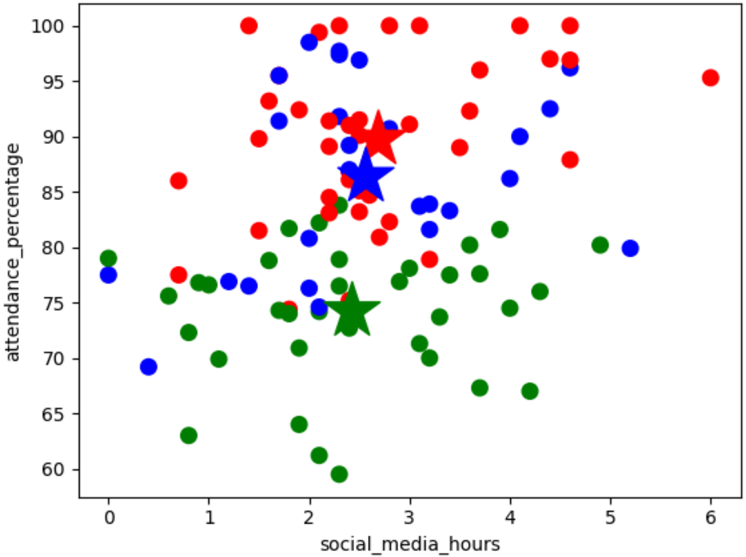

exam_score
90
80
70
60
50
40
30
20
0
1
2
3
4
5
6
7
study_hours_per_day
age
24
23
22
21
20
19
18

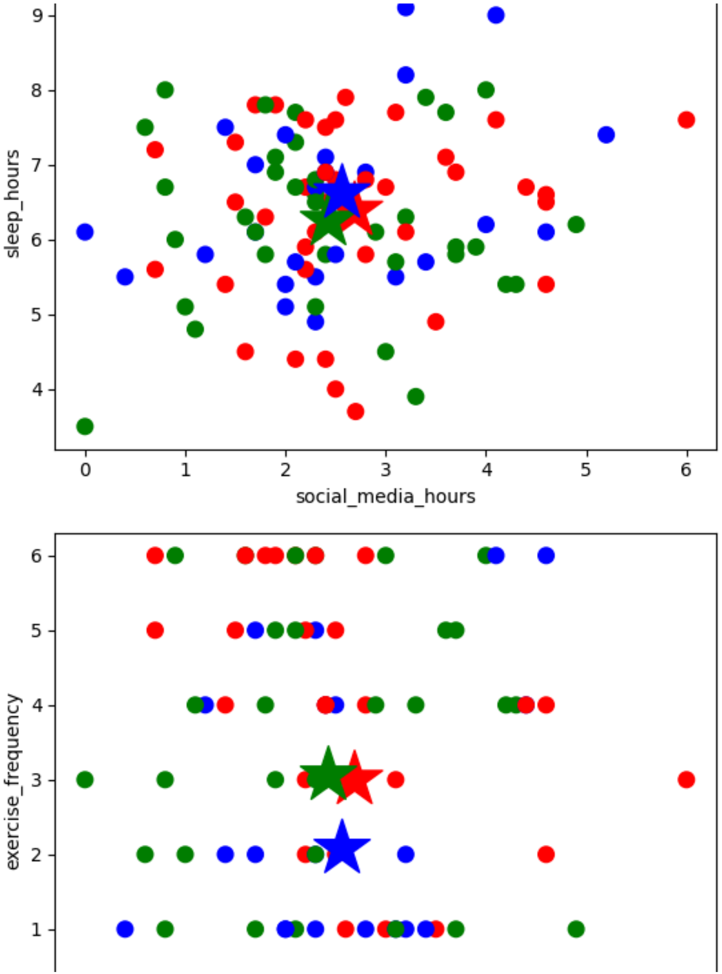

17
0
1
2
3
4
5
6
social_media_hours


7
6
5
4
3
2
1
0
study_hours_per_day
0
1
2
3
4
5
6
social_media_hours

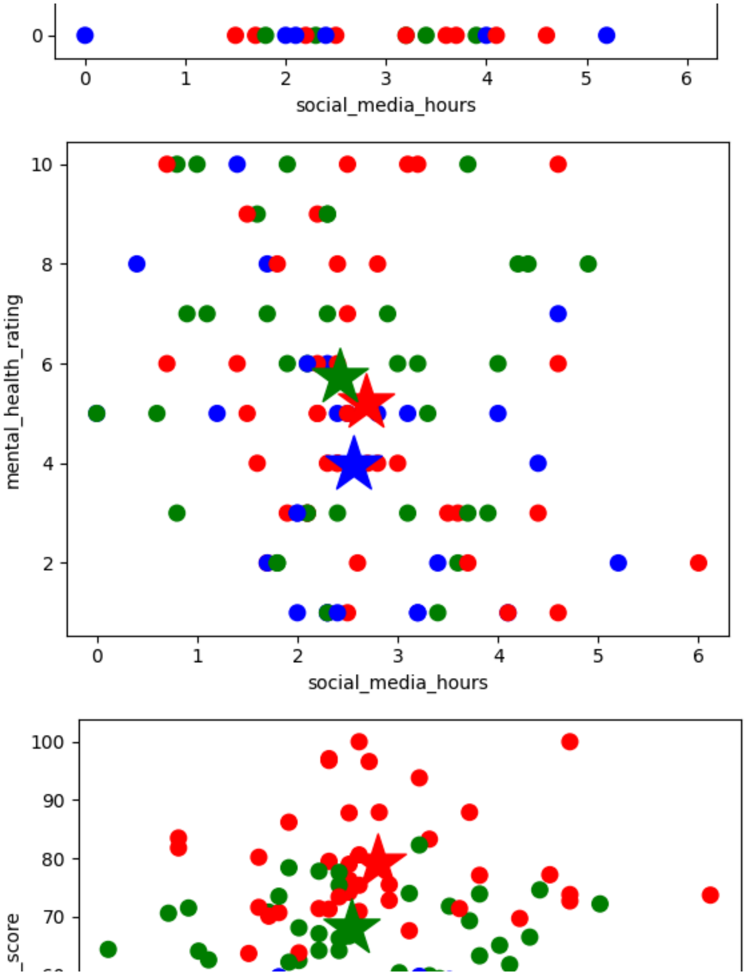

4.0
3.5
3.0
2.5
ours

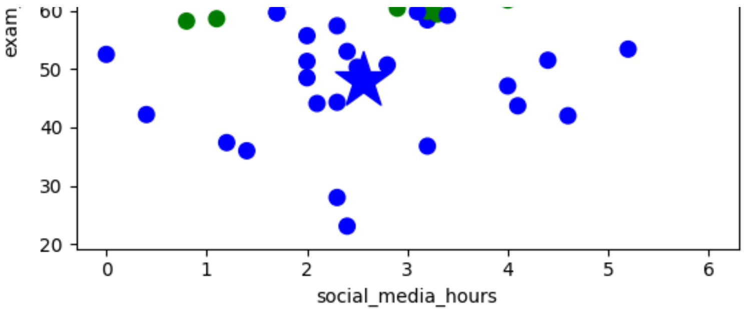

netflix_h
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
social_media_hours

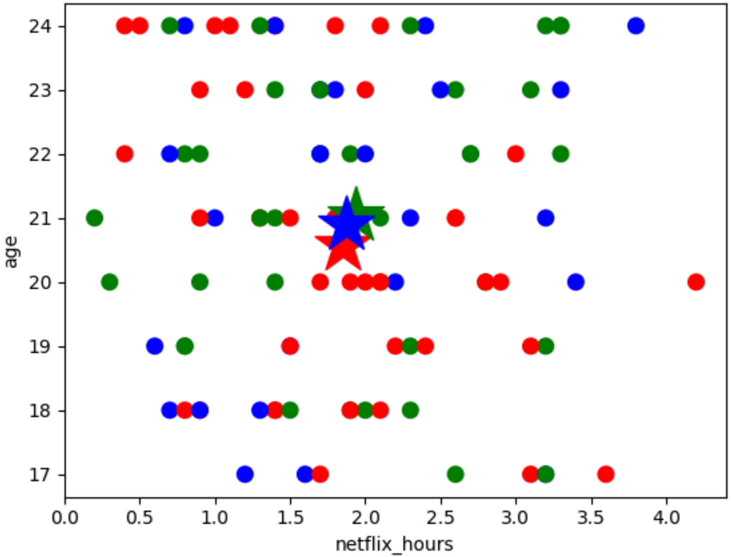

attendance_percentage
100
95
90
85
80
75
70
65
60
0
1
2
3
4
5
6
social_media_hours

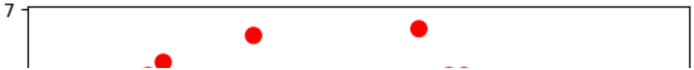


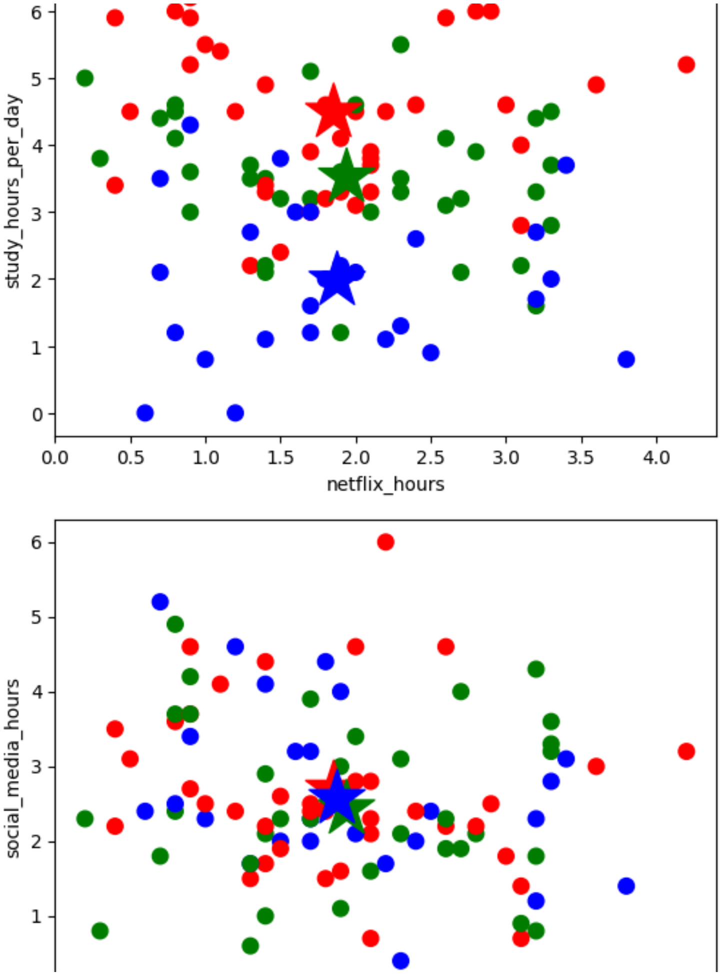
sleep_hours
9
8
7
6
5
4
0
1
2
3
4
5
6
social_media_hours
exercise_frequency
6
5
4
3
2
1


0
0
1
2
3
4
5
6
social_media_hours
mental_health_rating
10
8
6
4
2
0
1
2
3
4
5
6
social_media_hours
100
90
80
70
60
score

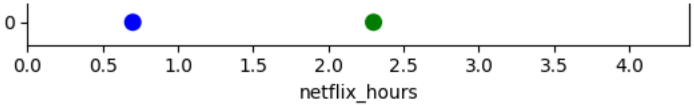

exam
60
50
40
30
20
0
1
2
3
4
5
6
social_media_hours

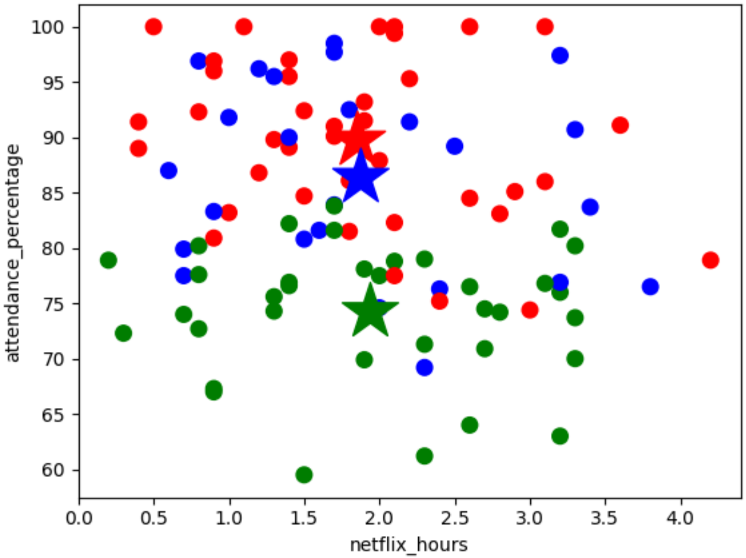

24
23
22
21
20
19
18
17
age
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours

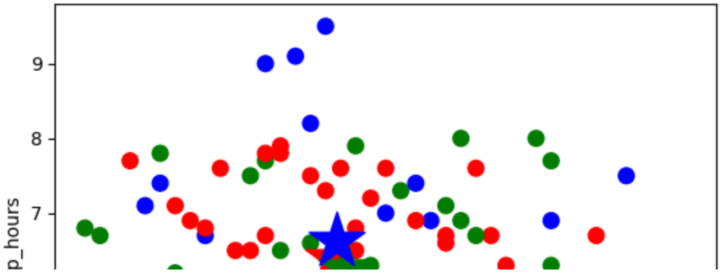

7

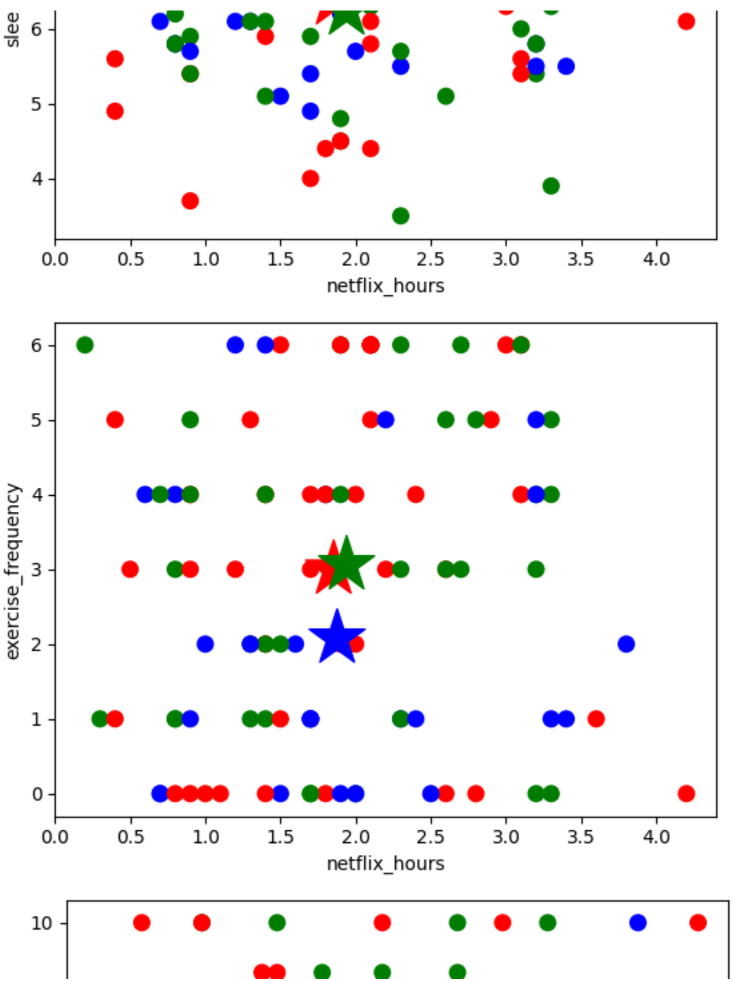


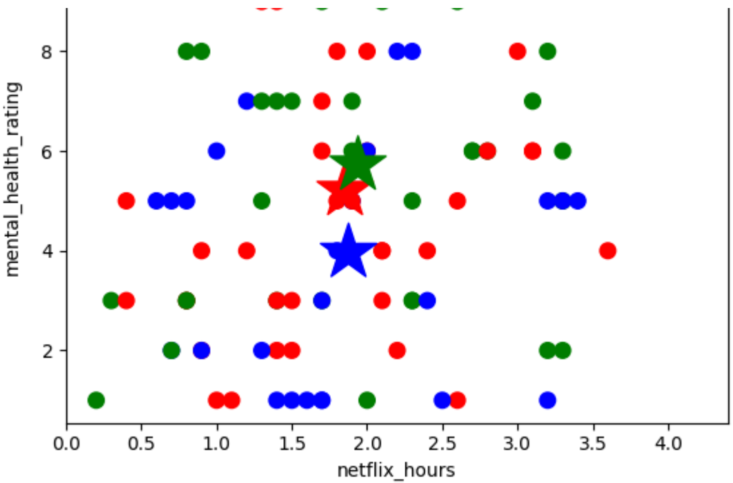
6
5
4
3
2
1
0
study_hours_per_day
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours
6
5
4
3
2
1
social_media_hours

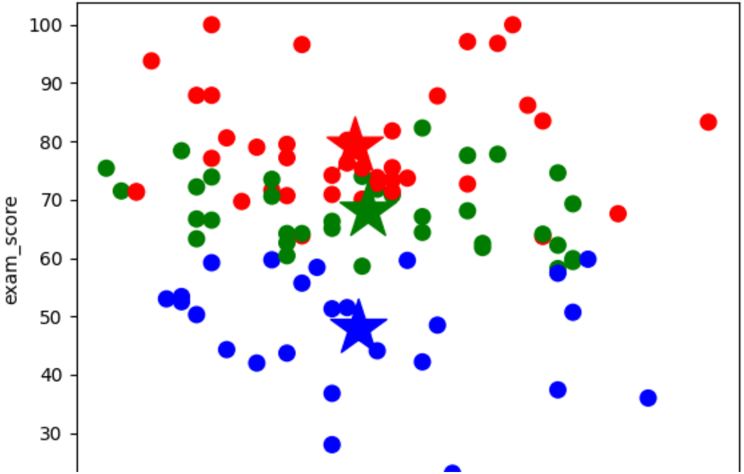

0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours

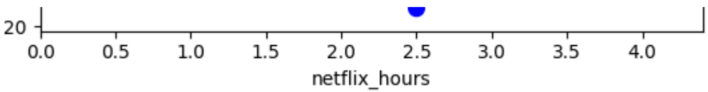

attendance_percentage
100
95
90
85
80
75
70
65
60
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours

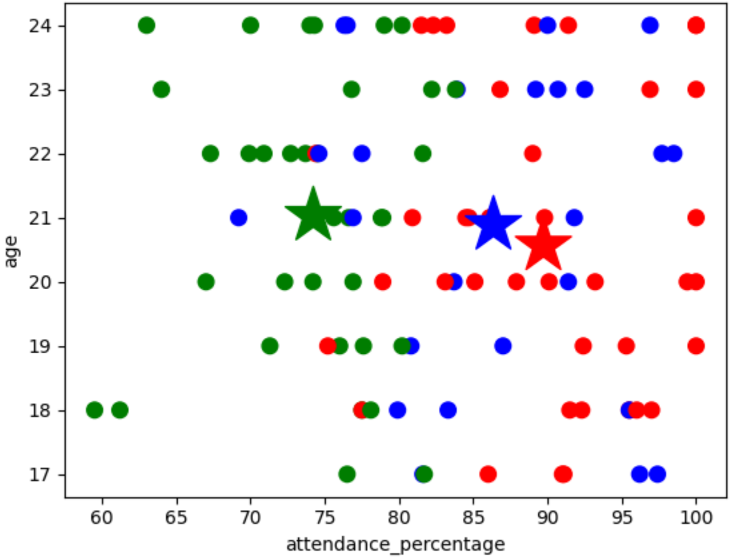

p_hours
9
8
7

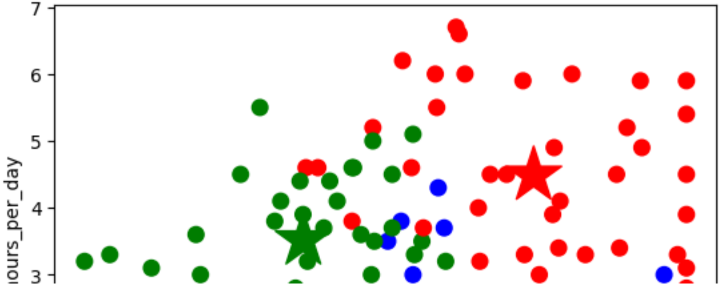

slee
6
5
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours
exercise_frequency
6
5
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours
10

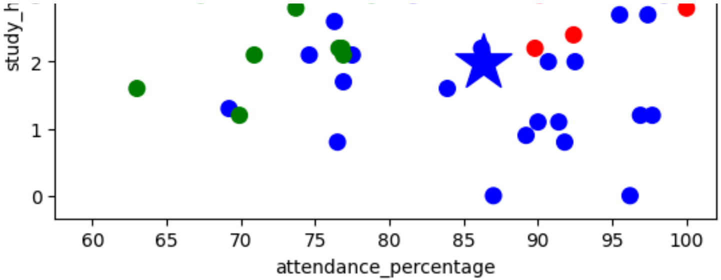

mental_health_rating
8
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours

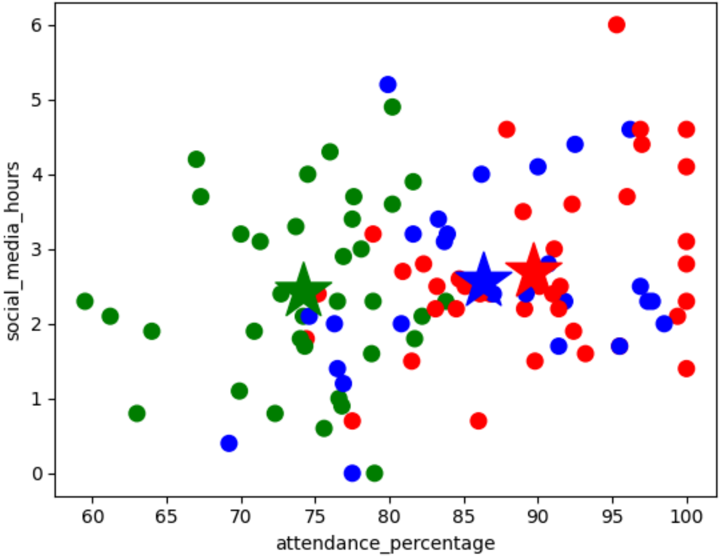

exam_score
100
90
80
70
60
50
40
30

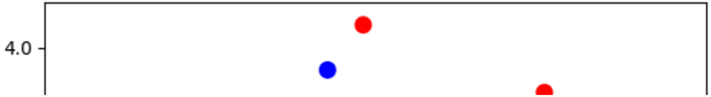

20
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
netflix_hours

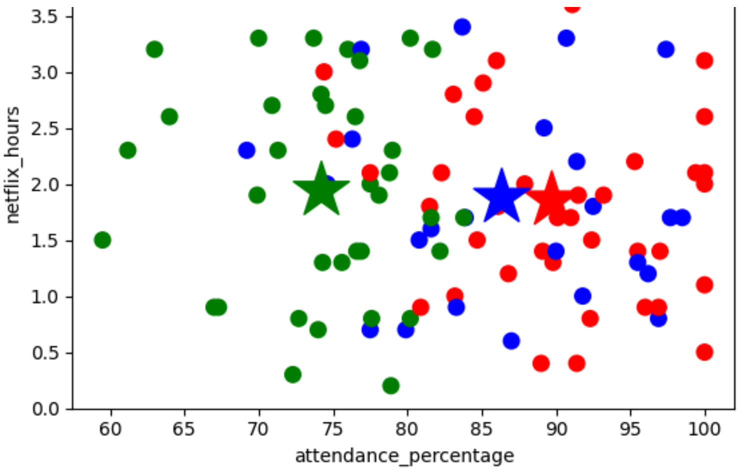

24
23
22
21
20
19
18
17
age
60
65
70
75
80
85
90
95
100
attendance_percentage

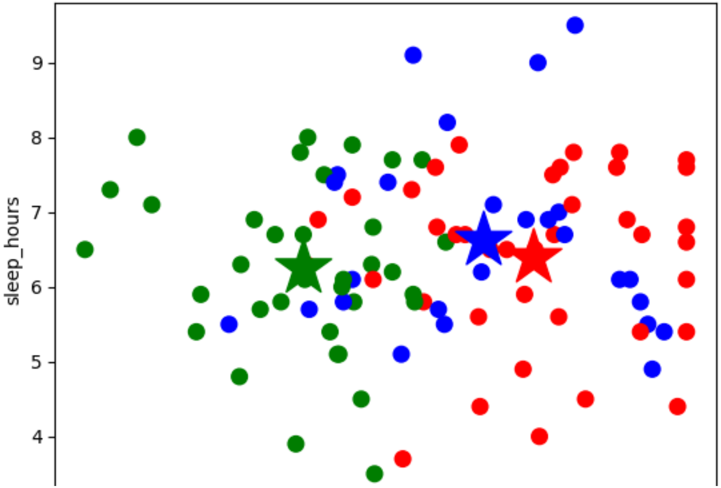

7
6
5
4
3
ours_per_day

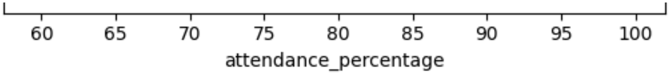

study_h
0
1
2
60
65
70
75
80
85
90
95
100
attendance_percentage

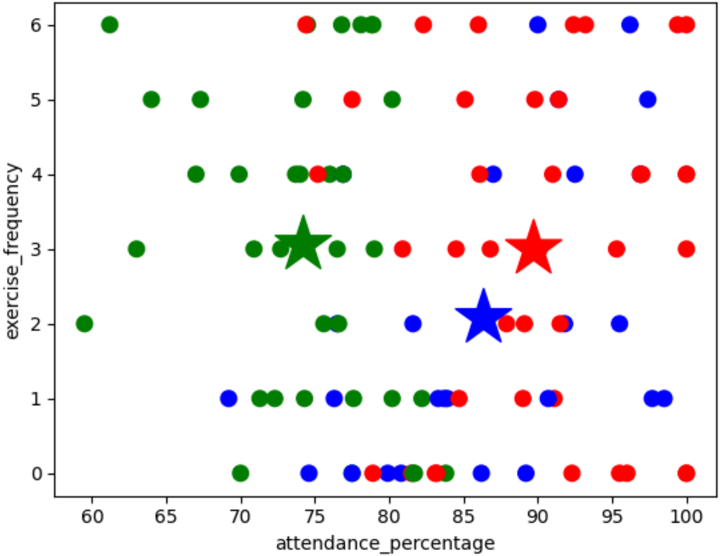

social_media_hours
0
1
2
3
4
5
6
60
65
70
75
80
85
90
95
100
attendance_percentage

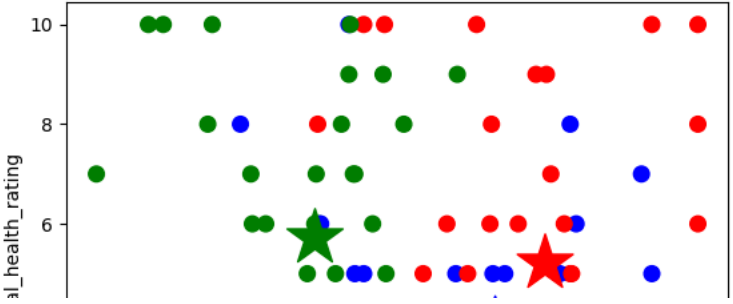

4.0

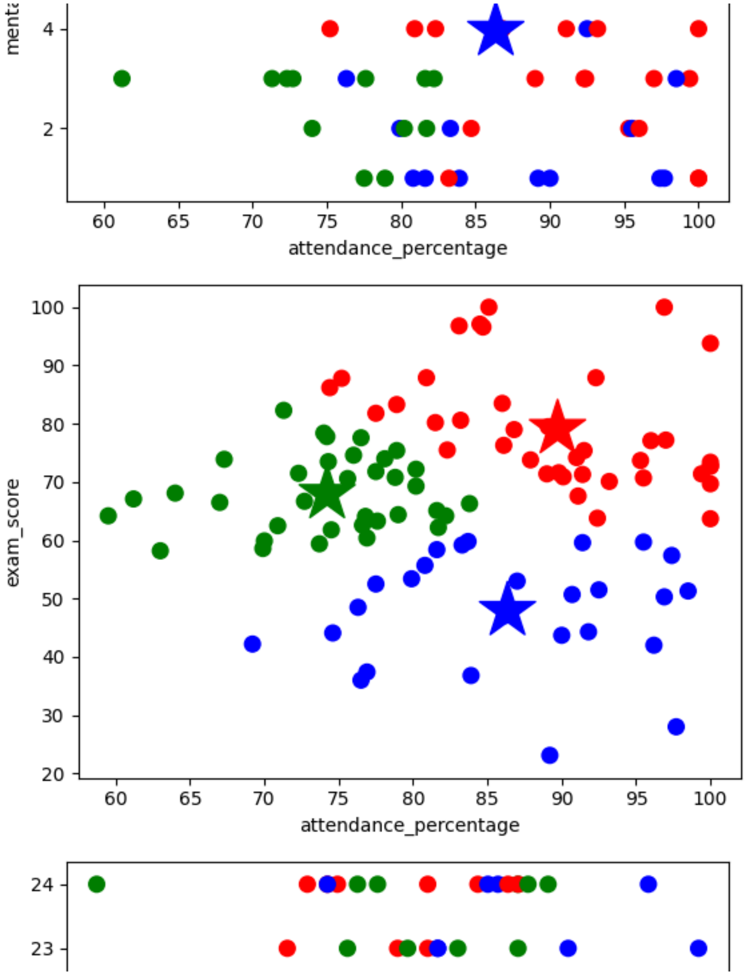

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
netflix_hours
60
65
70
75
80
85
90
95
100
attendance_percentage


9
8
7
6
5
4
sleep_hours

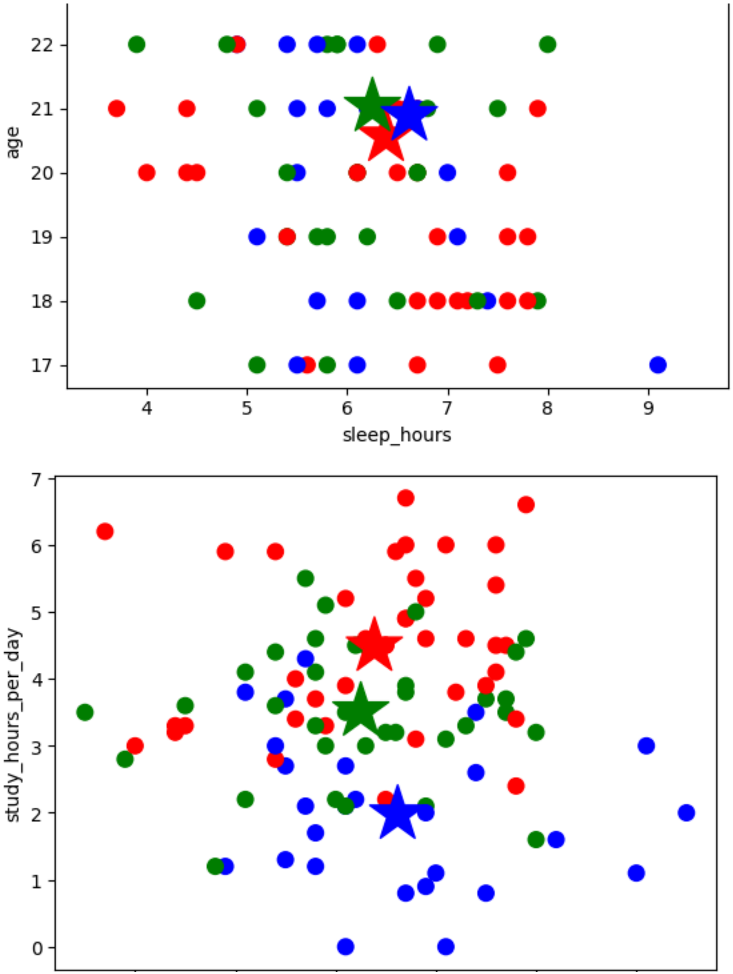

60
65
70
75
80
85
90
95
100
attendance_percentage

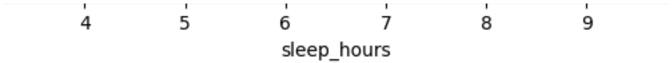

exercise_frequency
0
1
2
3
4
5
6
60
65
70
75
80
85
90
95
100
attendance_percentage

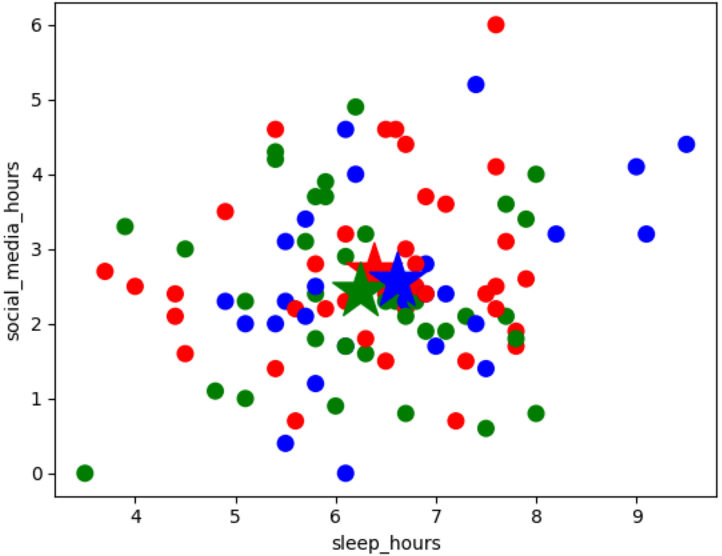

al_health_rating
6
8
10

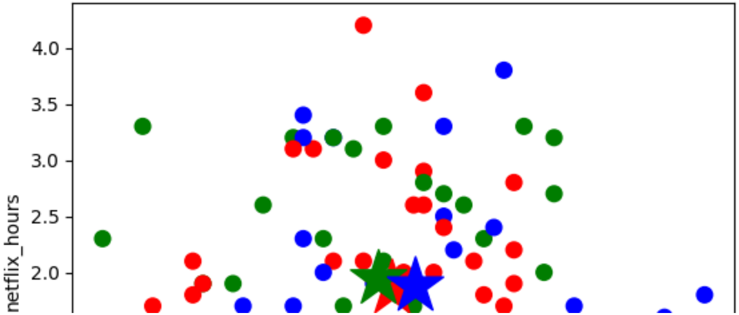


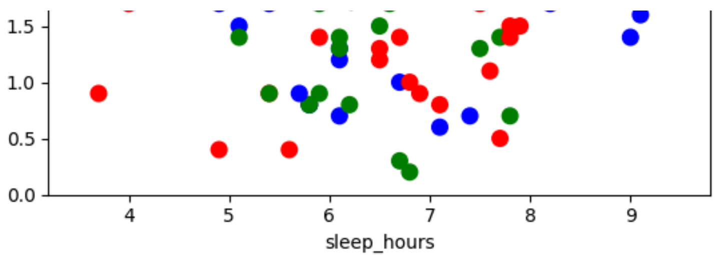
menta
4
2
60
65
70
75
80
85
90
95
100
attendance_percentage
exam_score
100
90
80
70
60
50
40
30
20
60
65
70
75
80
85
90
95
100
attendance_percentage
24
23

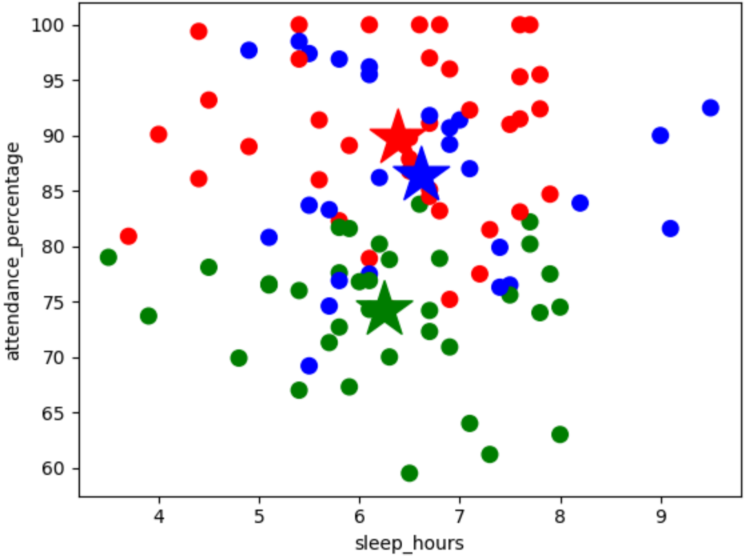

age
22
21
20
19
18
17
4
5
6
7
8
9
sleep_hours
study_hours_per_day
7
6
5
4
3
2
1
0

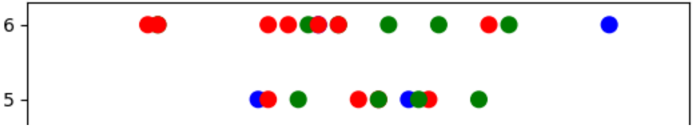

4
5
6
7
8
9
sleep_hours

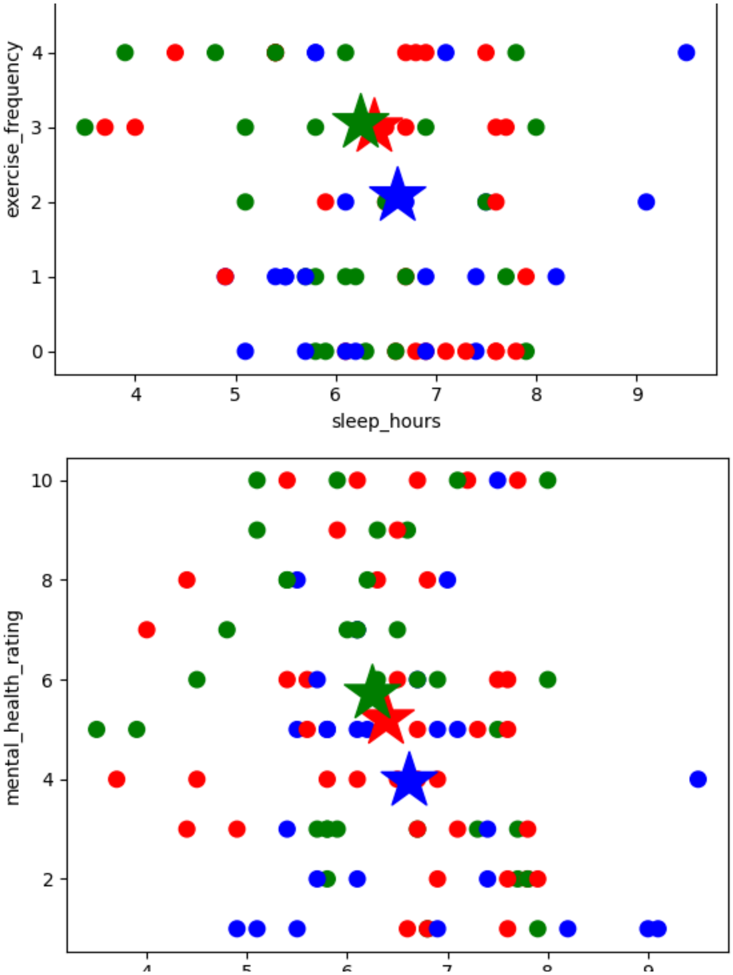

6
5
4
3
2
1
0
social_media_hours
4
5
6
7
8
9
sleep_hours


4.0
3.5
3.0
2.5
2.0
netflix_hours

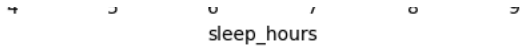

1.5
1.0
0.5
0.0
4
5
6
7
8
9
sleep_hours

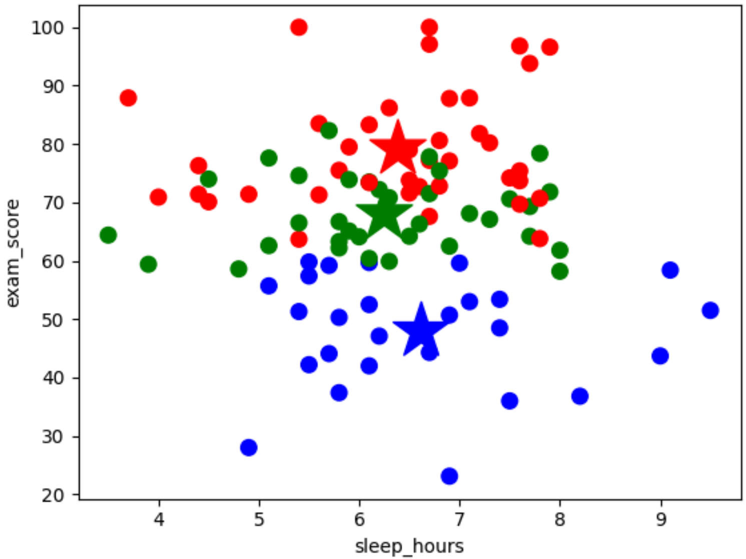

100
95
90
85
80
75
70
65
60
attendance_percentage
4
5
6
7
8
9
sleep_hours

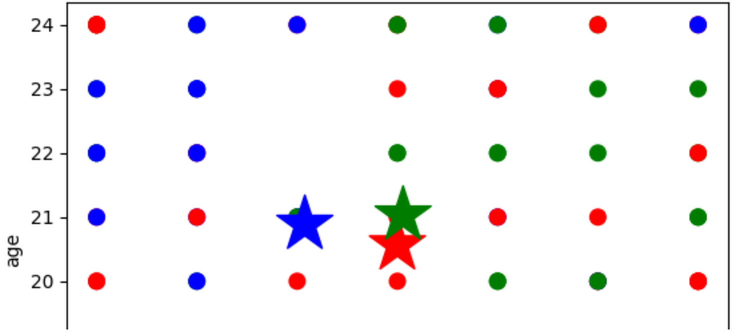

6
5

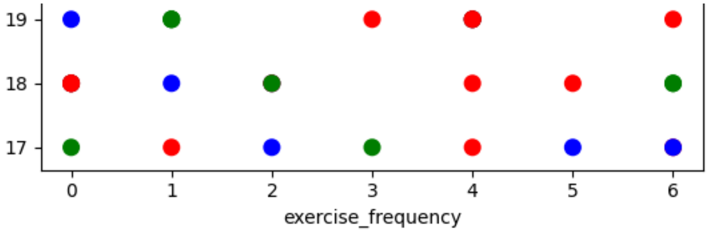


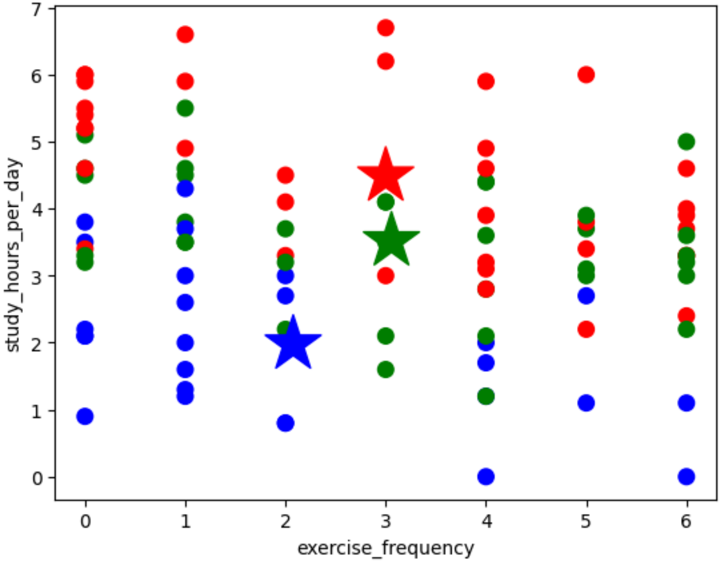
exercise_frequency
0
1
2
3
4
4
5
6
7
8
9
sleep_hours
mental_health_rating
2
4
6
8
10
4
5
6
7
8
9

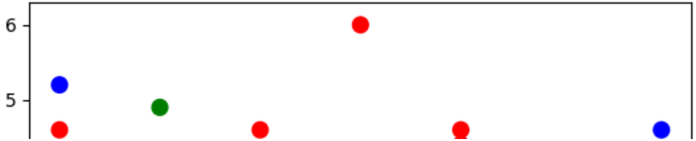

4
5
6
7
8
9
sleep_hours

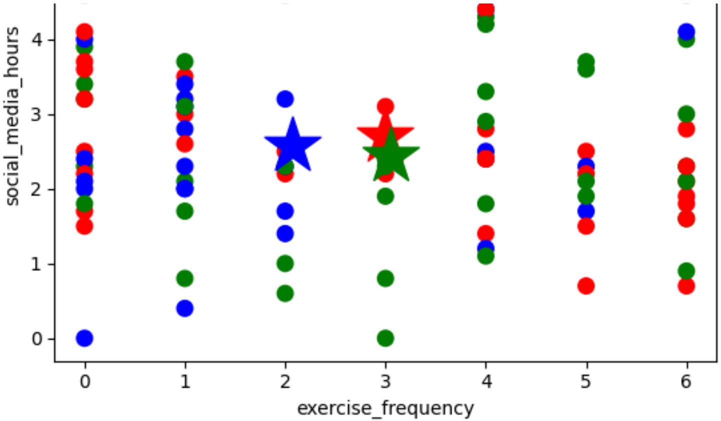

100
90
80
70
60
50
40
30
20
exam_score
4
5
6
7
8
9
sleep_hours

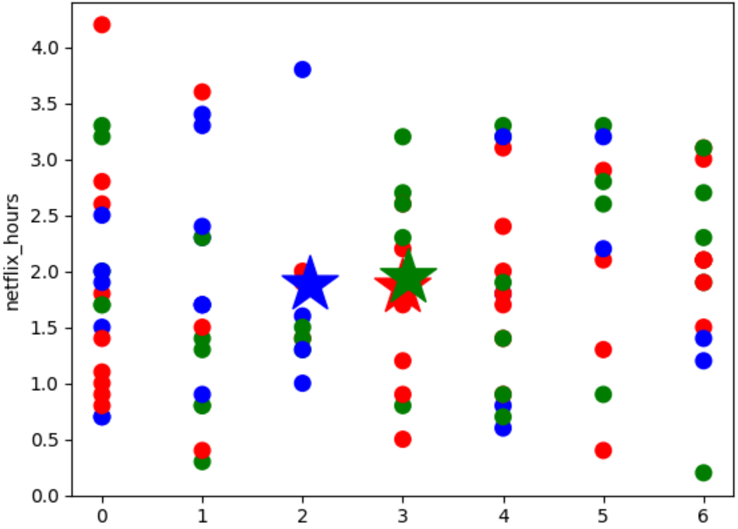

24
23
22
21
20
age

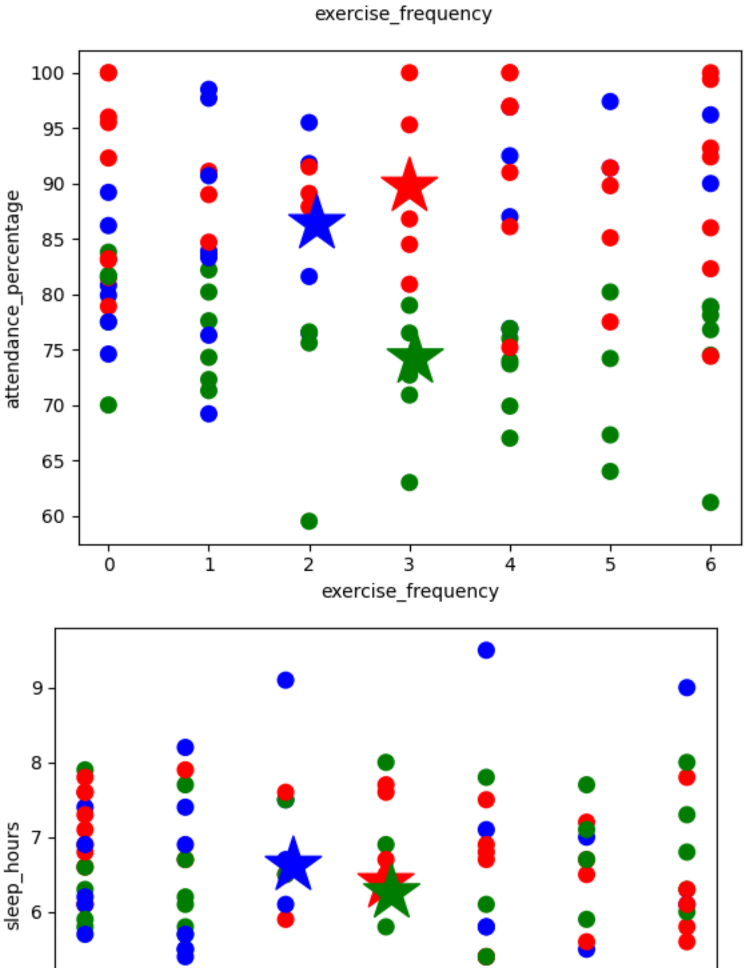

19
18
17
0
1
2
3
4
5
6
exercise_frequency

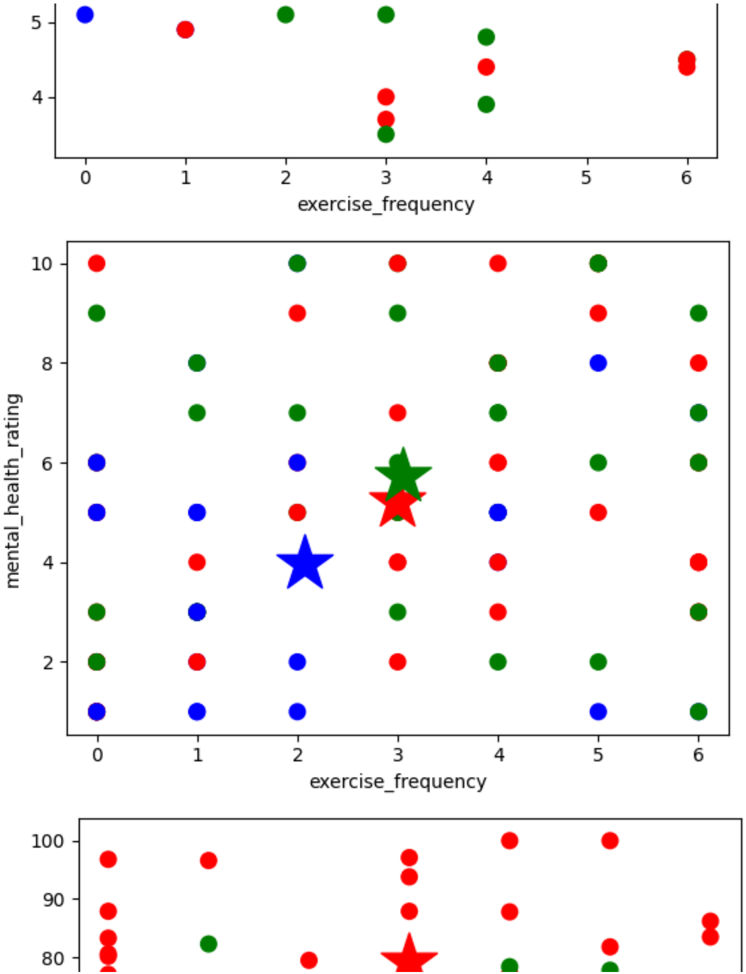

7
6
5
4
3
2
1
0
study_hours_per_day
0
1
2
3
4
5
6
exercise_frequency

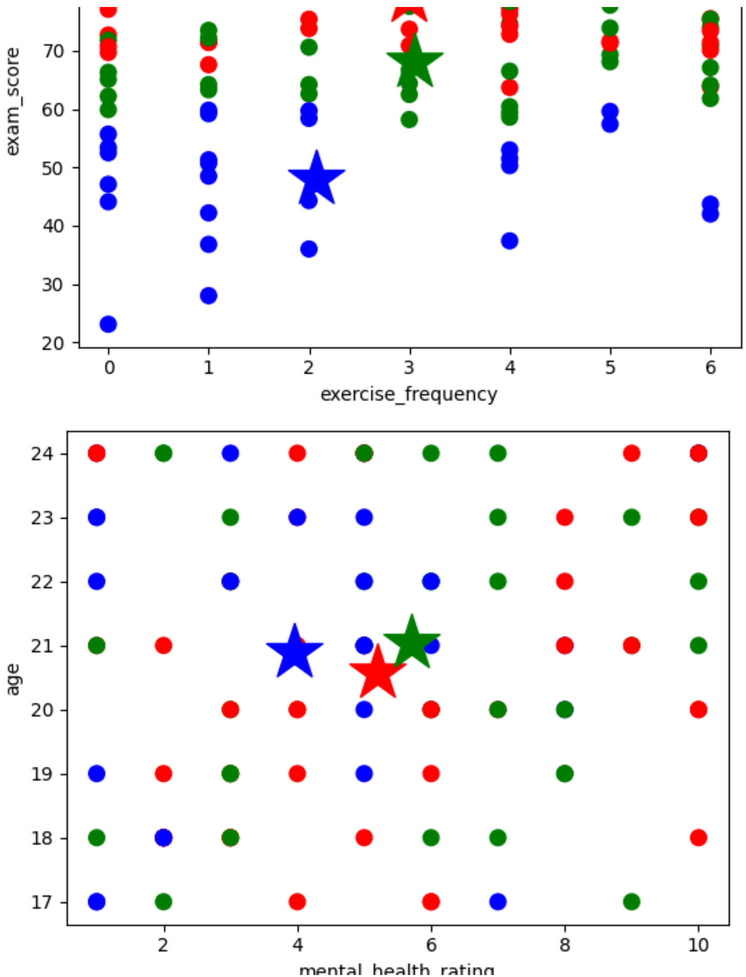

6
5

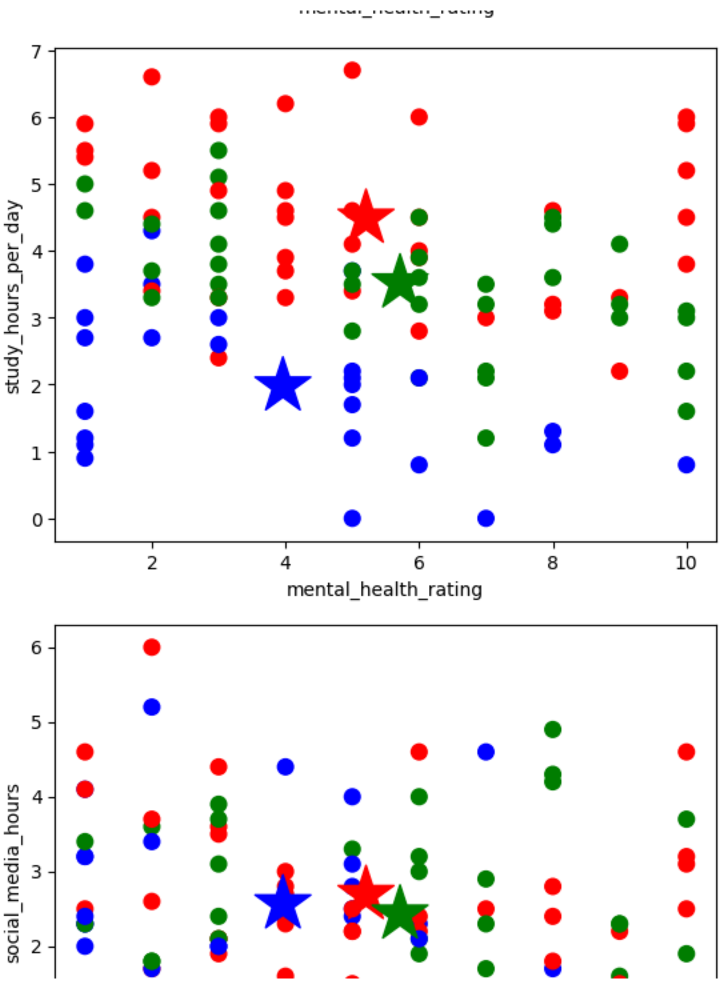

social_media_hours
0
1
2
3
4
0
1
2
3
4
5
6
exercise_frequency

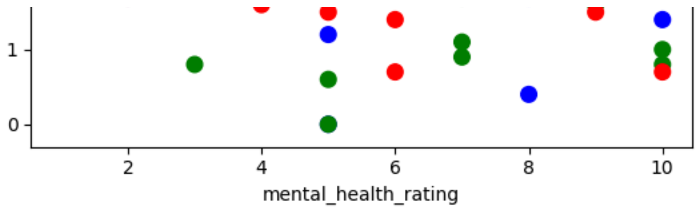

netflix_hours
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
1
2
3
4
5
6

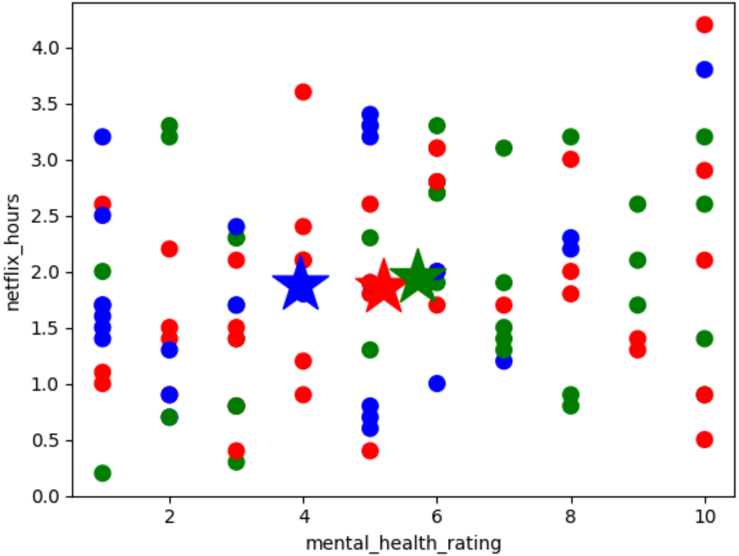

exercise_frequency
attendance_percentage
100
95
90
85
80
75
70
65
60
0
1
2
3
4
5
6
exercise_frequency
sleep_hours
9
8
7
6

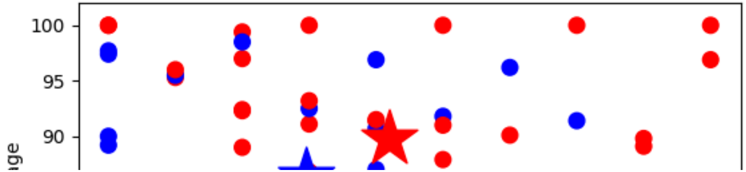

5
4
0
1
2
3
4
5
6
exercise_frequency
10
8
6
4
2
mental_health_rating
0
1
2
3
4
5
6
exercise_frequency
100
90
80

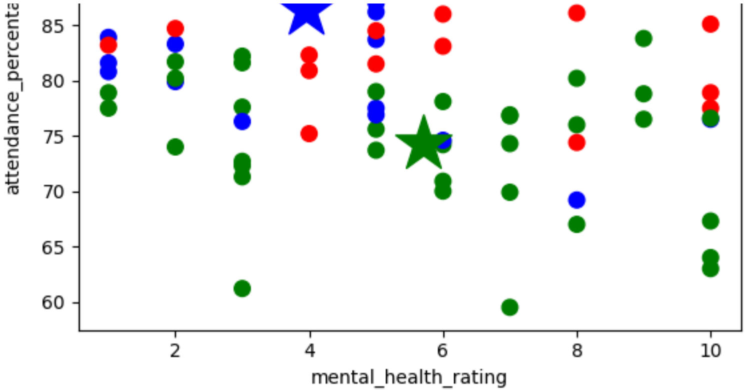

exam_score
70
60
50
40
30
20
0
1
2
3
4
5
6
exercise_frequency
age
24
23
22
21
20
19
18
17
2
4
6
8
10
mental_health_rating

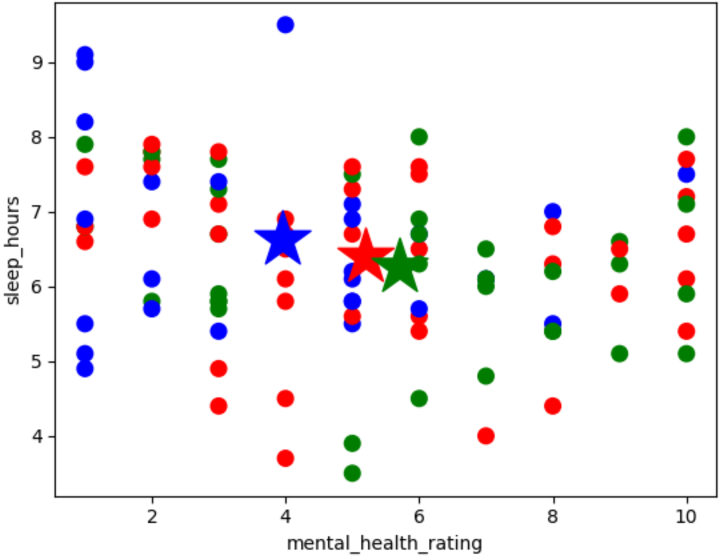

mental_health_rating
study_hours_per_day
0
1
2
3
4
5
6
7
2
4
6
8
10
mental_health_rating
social_media_hours
2
3
4
5
6

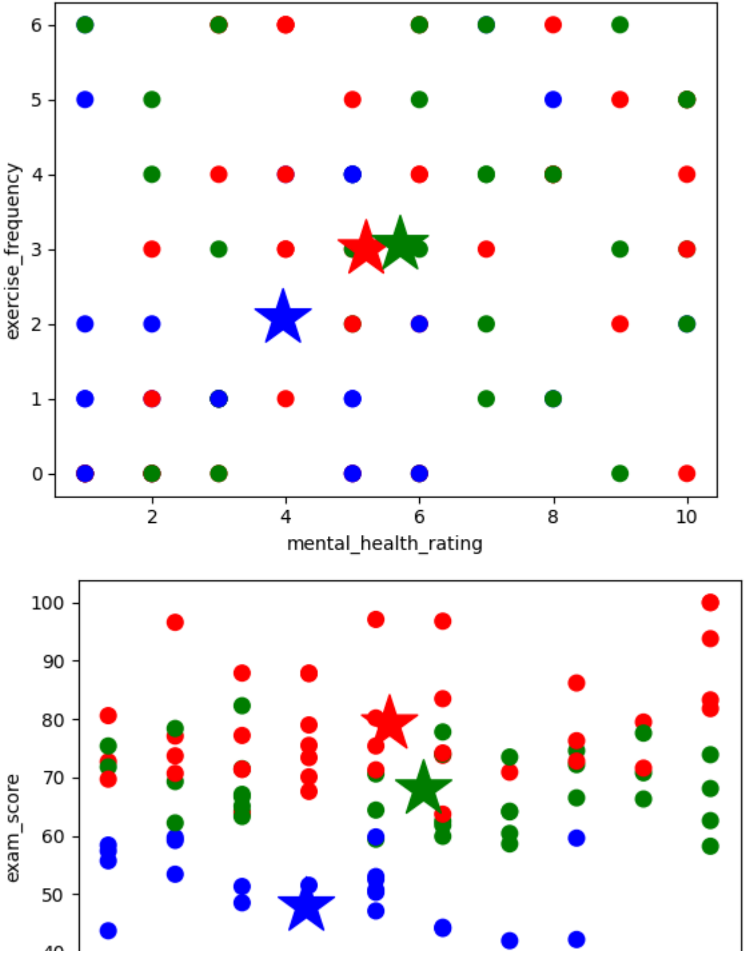

1
0
2
4
6
8
10
mental_health_rating


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
netflix_hours
2
4
6
8
10
mental_health_rating

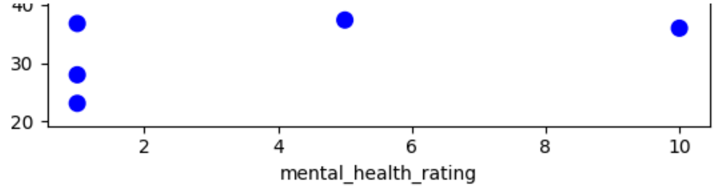

100
95
90
ge

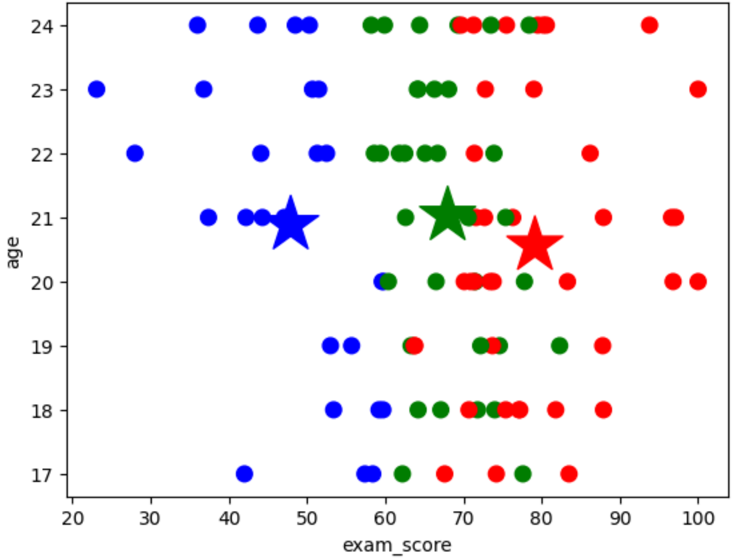

attendance_percenta
85
80
75
70
65
60
2
4
6
8
10
mental_health_rating

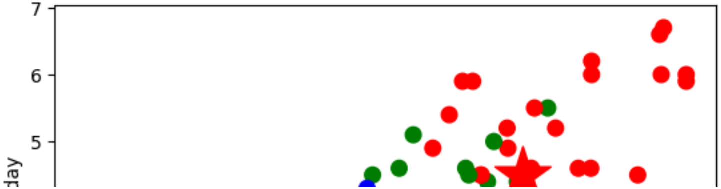

sleep_hours
9
8
7
6
5
4
2
4
6
8
10
mental_health_rating

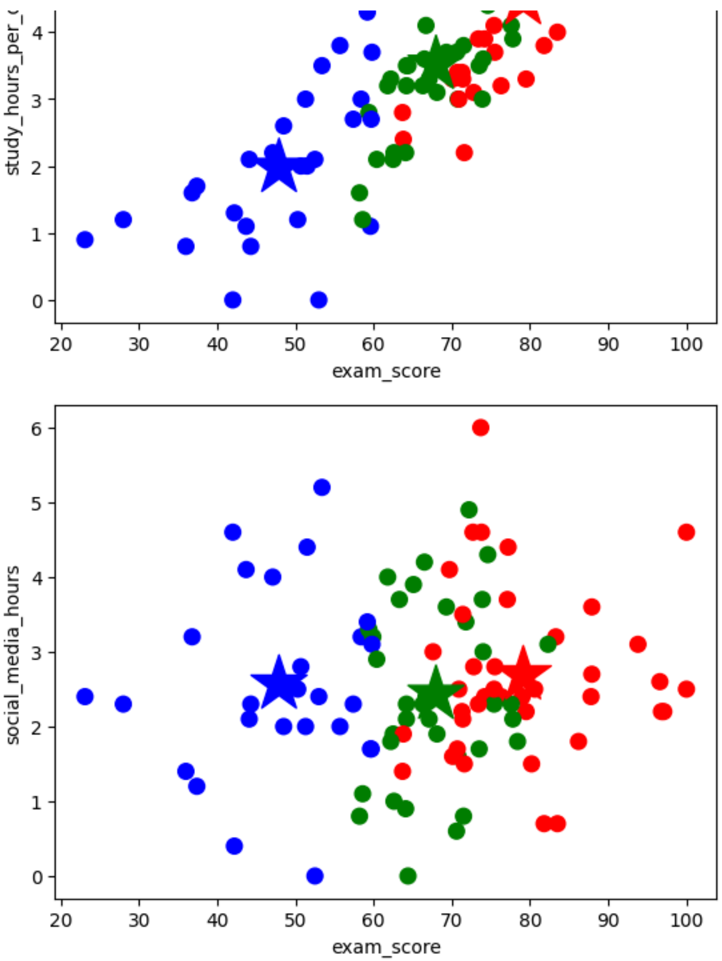


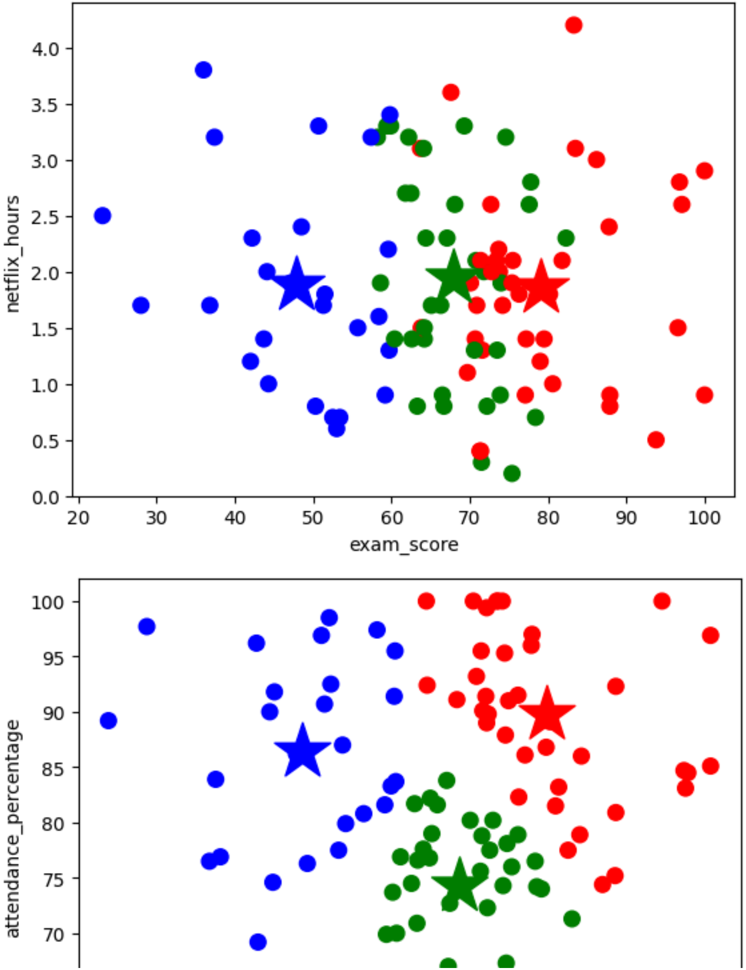
exercise_frequency
6
5
4
3
2
1
0
2
4
6
8
10
mental_health_rating
exam_score
100
90
80
70
60
50
40


40
30
20
2
4
6
8
10
mental_health_rating


age
24
23
22
21
20
19
18
17
20
30
40
50
60
70
80
90
100
exam_score


7
6
5
day


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


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
netflix_hours
20
30
40
50
60
70
80
90
100
exam_score
100
95
90
85
80
75
70
attendance_percentage

65