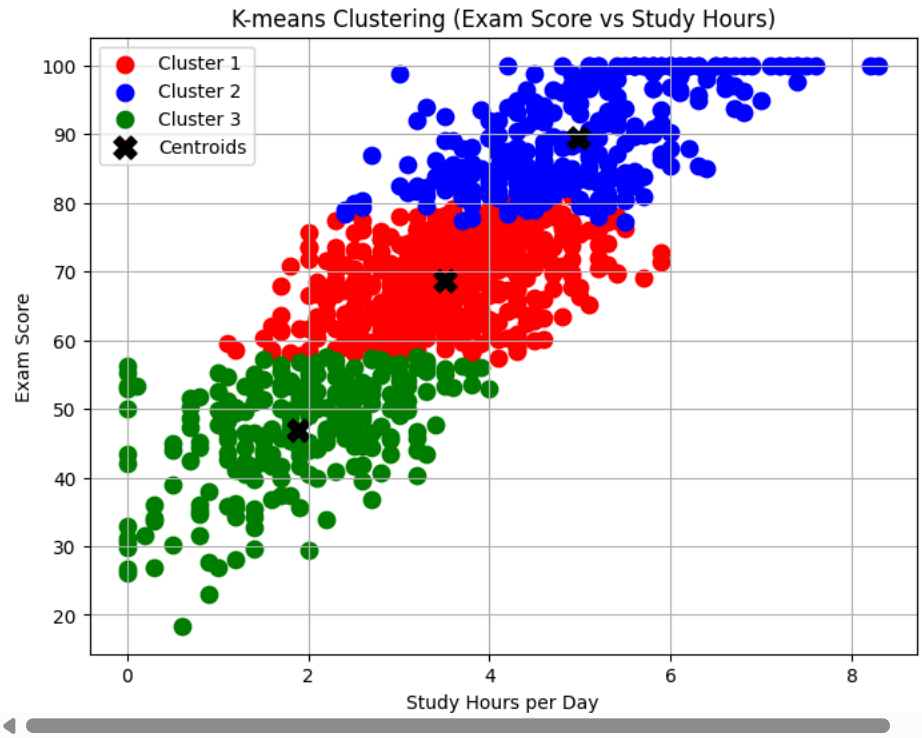
```
from google.colab import files
uploaded = files.upload()

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans


df = pd.read_csv("student_habits_performance.csv")

# - 'student_id' is an identifier
# - Categorical variables are not used here because they require special encoding (outside current scope)

numeric_data = df.select_dtypes(include=[np.number])

data_points = numeric_data.to_numpy()

K = 3 #The value of k was set to 4 based on visual inspection of the scatter plot and the assumption that students can be grouped into fo

# use scikit-learn to calculate the centroids
kmeans = KMeans(n_clusters=K, random_state=0, n_init=10)
labels = kmeans.fit_predict(data_points)
centroids = kmeans.cluster_centers_

plt.figure(figsize=(8, 6))
colors = ['red', 'blue', 'green', 'purple']
for k in range(K):
    cluster = data_points[labels == k]
    plt.scatter(cluster[:, 1], cluster[:, -1], color=colors[k], label=f'Cluster {k+1}', s=80)

plt.scatter(centroids[:, 1], centroids[:, -1], color='black', marker='X', s=150, label='Centroids')
plt.xlabel('Study Hours per Day')
plt.ylabel('Exam Score')
plt.title('K-means Clustering (Exam Score vs Study Hours)')
plt.legend()
plt.grid(True)
plt.show()

```

Elegir archivos student_hab…rmance.csv

- **student_habits_performance.csv**(text/csv) - 73663 bytes, last modified: 6/5/2025 - 100% done

```
Saving student_habits_performance.csv to student_habits_performance (10).csv
```



```
1. Do you think these centers could be representative of the data? Why

Yes, they seem quite representative. Each centroid is located near the
of its respective data set.

2. How do you get the value of k for use?

The value of k = 3 may have been chosen after a visual inspection of t
with 3 groups, the data are divided more clearly.

Visual consistency and algorithm behavior showed that k = 3 avoids unn
overlaps between groups.

3. Would the centers use more if they used more? Less value?

*   If k is larger (such as k = 4), some groups overlap, making interp
difficult. Unnecessary clusters are formed that provide no real value.
*   If k were smaller (such as k = 2), this shows a natural separation
overall performance.

So there is a trade-off: higher means more granular, which means more
The corporeality depends on the actual data structure.

4. How far apart are the centers? Are any close to others?

*   Poor performance and few hours of study.
*   Average performance with regular study habits.
*   High performance and many hours of study.

5. What would happen to the centers if we had many outliers in the
box-and-whisker analysis?

Outliers would pull the centroids toward extreme values, potentially m
the center of the true cluster.

6. What can you say about the data based on the centers?

* Students who study little and have low scores.
* Students with moderate habits and average performance.
* Students who study more and tend to have high scores.

https://github.com/SvpremeLemus/Act
```

1. Do you think these centers could be representative of the data? Why?

Yes, they seem quite representative. Each centroid is located near the center of its respective data set.

2. How do you get the value of k for use?

The value of k = 3 may have been chosen after a visual inspection of the graph: with 3 groups, the data are divided more clearly.

Visual consistency and algorithm behavior showed that k = 3 avoids unnecessary overlaps between groups.

3. Would the centers use more if they used more? Less value?

- If k is larger (such as k = 4), some groups overlap, making interpretation difficult. Unnecessary clusters are formed that provide no real value.
- If k were smaller (such as k = 2), this shows a natural separation based on overall performance.

So there is a trade-off: higher means more granular, which means more general. The corporeality depends on the actual data structure.

4. How far apart are the centers? Are any close to others?

- Poor performance and few hours of study.
- Average performance with regular study habits.
- High performance and many hours of study.

5. What would happen to the centers if we had many outliers in the box-and-whisker analysis?

Outliers would pull the centroids toward extreme values, potentially misplacing the center of the true cluster.

6. What can you say about the data based on the centers?

- Students who study little and have low scores.
- Students with moderate habits and average performance.
- Students who study more and tend to have high scores.

https://github.com/SvpremeLemus/Act