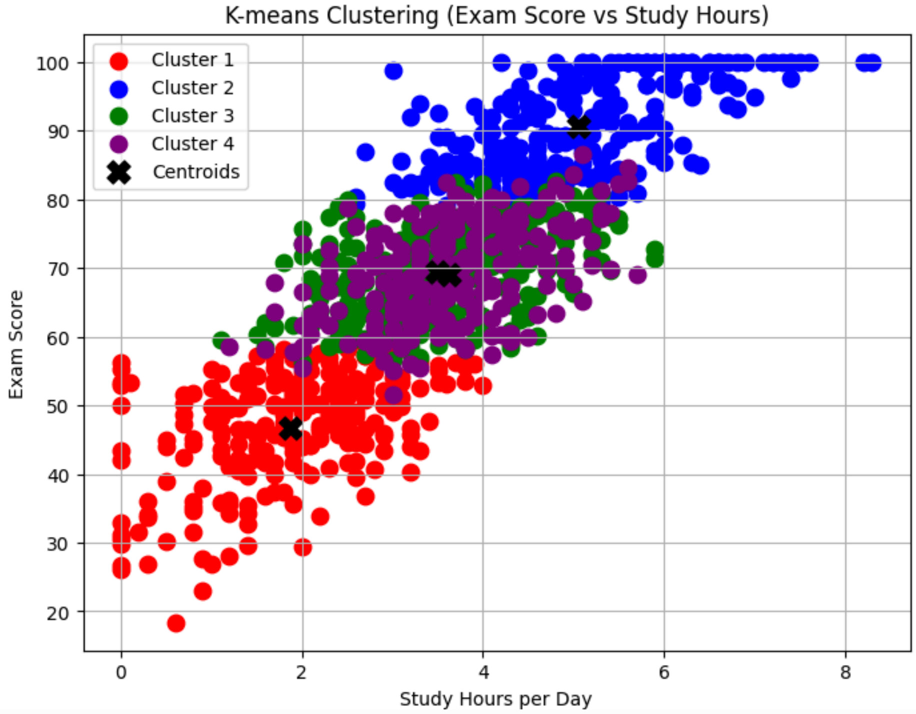
```python
from google.colab import files
uploaded = files.upload()

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 = 4 #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 (7).csv
```



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

Indeed, they seem quite representative. Each of them is located around
center of its respective group of data points, showing that the algori
successfully grouped students with similar characteristics and test sc

2. How do you get the value of k for use?
The value of k = 4 was likely based on the decisions made;

*   Visual inspection of data distribution.
*   Prior knowledge of how many natural clusters can be expected.
*   Alternatively, the elbow method is used to define where adding mor
clusters no longer significantly improves the fit.

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

*   Most of the value can be obtained in smaller, more connected group
can overwhelm the data and reduce interpretation.
*   A lower value of k can confuse different samples into larger group
lacking meaningful distinctions.


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?

From the plots;

*   Some centroids are well separated, as in the red and blue clusters
*   Others, especially green and purple, seem to be closer together, w
indicates the behavior of overlapping clusters with insufficient
differentiation.

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?

There are distinct groups of students based on their study hours and e
performance. Students who study more tend to score higher, as shown by
upward trend in the plot.
Some students may study less but still perform well, possibly due to o
factors (like natural aptitude).

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

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

Indeed, they seem quite representative. Each of them is located around the center of its respective group of data points, showing that the algorithm successfully grouped students with similar characteristics and test scores.

2. How do you get the value of k for use? The value of k = 4 was likely based on the decisions made;

- Visual inspection of data distribution.
- Prior knowledge of how many natural clusters can be expected.
- Alternatively, the elbow method is used to define where adding more clusters no longer significantly improves the fit.

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

- Most of the value can be obtained in smaller, more connected groups, which can overwhelm the data and reduce interpretation.
- A lower value of k can confuse different samples into larger groups, lacking meaningful distinctions.

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?

From the plots;

- Some centroids are well separated, as in the red and blue clusters.
- Others, especially green and purple, seem to be closer together, which indicates the behavior of overlapping clusters with insufficient differentiation.

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?

There are distinct groups of students based on their study hours and exam performance. Students who study more tend to score higher, as shown by the upward trend in the plot. Some students may study less but still perform well, possibly due to other factors (like natural aptitude).

https://github.com/SvpremeLemus/Act