**1. Data Preprocessing Techniques: My Choices**

To get the data ready for analysis, I implemented the following preprocessing steps, and here's why I chose each technique:

* **Data Cleaning (Handling Missing Values):** I noticed some missing data in the dataset. To deal with this, I decided to use **mean imputation** for the numerical columns. My reasoning was that using the average value would help maintain the overall distribution of the data without losing too many data points. For the categorical columns, I opted for **mode imputation**, replacing the missing values with the most frequent category. This seemed like the most sensible approach to fill in missing categorical information without introducing arbitrary values.
* **Data Integration:** Since I was working with a single dataset, **data integration wasn't necessary** for this analysis. If I had wanted to combine this data with other sources in the future, I would have needed to consider how to merge the datasets based on common identifiers.
* **Data Reduction (Attribute Subset Selection):** To make the cluster analysis more focused and interpretable, I chose to work with a subset of the available columns. I selected math\_score, reading\_score, writing\_score, and gaming\_hours. My thinking was that these features are the most directly relevant to understanding how academic success relates to gaming habits. Including less pertinent features might have added noise to the clustering process.
* **Data Transformation (Normalization):** Because the numerical features I selected have different scales (scores might range from 0-100, while gaming hours could be 0-24), I applied **Min-Max scaling**. This method scales all the values to a range between 0 and 1. I chose this because it ensures that no single feature dominates the distance calculations in the K-Means algorithm simply due to its larger numerical range. This way, each feature contributes more equally to the clustering.
* **Data Discretization:** For this particular K-Means clustering task, I didn't directly discretize any of the features. However, I considered that for other types of analysis or visualization, it might be useful to group the gaming\_hours into categories (like "low," "medium," "high"). For now, I kept it as a continuous numerical variable for the clustering algorithm.

**2. Choice of Model: Cluster Analysis (K-Means) and Why It Fits**

For this analysis, I decided to use **Cluster Analysis**, specifically the **K-Means algorithm**. My goal was to see if there were natural groupings of students based on their academic performance and how much time they spend playing games, and K-Means is a well-established method for this kind of task.

**Here's why I felt K-Means was a good fit for this dataset:**

* **Unsupervised Learning:** I didn't have any pre-existing labels or categories for the students. I wanted to discover inherent structures in the data itself, and K-Means is an unsupervised learning technique that excels at finding these hidden patterns.
* **Segmentation Goal:** My aim was to segment the student population into distinct groups with similar characteristics regarding their academic achievements and their level of engagement with gaming. K-Means is designed to partition data points into clusters based on their similarity.
* **Interpretability of Results:** I wanted to be able to understand the characteristics of the different student groups that emerged. K-Means produces clusters with centroids (mean values), which makes it relatively straightforward to interpret the typical academic scores and gaming hours for each group.
* **Suitability for Numerical Data:** The key features I selected for clustering (math\_score, reading\_score, writing\_score, gaming\_hours) are numerical. K-Means works effectively with numerical data, especially after I normalized the scales.
* **Exploratory Analysis:** This was primarily an exploratory analysis to see what kind of student segments might exist in the data. K-Means is a popular and efficient algorithm for this type of initial investigation.

To help me determine a good number of clusters (K) for the K-Means algorithm, I used the **Elbow Method**. By plotting the within-cluster sum of squares (inertia) for different values of K, I looked for a point where adding more clusters started to yield diminishing returns. Based on the elbow plot I generated, I chose K=3 as a potentially meaningful number of clusters for this dataset.

By applying K-Means with 3 clusters, I was able to assign each student to a specific group. By then looking at the average scores and gaming hours for each cluster, I could start to understand the different profiles of students within the dataset. For example, one cluster might represent students with high academic performance and low gaming hours, while another might show a different pattern. This kind of insight could be valuable for further research or for developing targeted strategies for different student groups.