**Introduction to Variable Categorization for EDA**

In order to understand the underlying patterns and factors that may influence a student’s academic outcome (Target: *Dropout*, *Graduate*, or *Enrolled*), it is essential to categorize the variables in the dataset into meaningful groups. These groups reflect different aspects of a student's profile and learning environment, and analyzing them separately enables a more focused and interpretable Exploratory Data Analysis (EDA).

Based on the variables available in the dataset, we can group them into five key dimensions:

1. **Academic Performance** – indicators of student grades, credits, and progress.
2. **Socioeconomic Background** – financial and social factors that may affect learning.
3. **Motivation Levels** – signs of student ambition or commitment at the time of applying.
4. **Attendance Patterns** – records that may reflect presence or participation in the program.
5. **Psychological or Personal Factors** – demographic and personal characteristics that might influence outcomes.

**1. Academic Performance**

These variables directly reflect a student’s academic achievements or evaluation results:

* avg\_grade – Average grade
* Admission grade – Grade upon admission
* total\_approved – Number of units approved
* total\_credited – Number of credits earned
* units\_with\_evaluation – Units that had evaluations
* units\_without\_evaluation – Units without evaluations
* parental\_scores – Scores based on parental feedback or influence

**💰 2. Socioeconomic Background**

These variables describe the financial and social context of the student:

* Scholarship holder – Whether the student receives a scholarship
* Tuition fees up to date – Whether tuition fees are fully paid
* Debtor – Whether the student is in debt
* GDP – GDP level at time of enrollment
* Unemployment rate – National unemployment rate
* Inflation rate – Inflation rate (may indicate economic pressure)
* Nacionality – Country of origin (proxy for access/opportunity)

**🔥 3. Motivation Levels**

These reflect the student's interest, enthusiasm, or decision-making at the time of application:

* Application mode – How the student applied
* Application order – Ranking or priority of the application
* Course – Chosen field of study (some courses may indicate intrinsic motivation)

**⏰ 4. Attendance Patterns**

These provide insight into participation and academic engagement:

* total\_enrolled – Number of units the student enrolled in
* Daytime/evening attendance – Mode of attendance (may reflect engagement level)

**🧠 5. Psychological/Personal Factors**

These variables may reflect personal challenges or demographic characteristics that influence academic performance:

* Age at enrollment – Age when student enrolled
* Gender – Male (1) or Female (0)
* Displaced – Whether the student is displaced (e.g., refugee or similar condition)
* International – Whether the student is international
* Educational special needs – Whether the student has special learning needs
* Marital status – May impact time and emotional availability

Data:

Composites variables:

**Summary of Calculations for Key Academic Metrics**

In the **academic5** dataset, we performed several calculations to generate important academic metrics. These new variables provide a clearer picture of student academic progress across different semesters and subjects. Below is a breakdown of each metric and its calculation:

1. **total\_enrolled**:
   * **Definition**: This represents the total number of curricular units a student has enrolled in, combining both the 1st and 2nd semesters.
   * **Calculation**:

total\_enrolled=Curricular units 1st sem (enrolled)+Curricular units 2nd sem (enrolled)\

1. **total\_approved**:
   * **Definition**: This metric indicates the total number of curricular units a student has approved (passed) across both semesters.
   * **Calculation**:

total\_approved=Curricular units 1st sem (approved)+Curricular units 2nd sem (approved)

1. **avg\_grade**:
   * **Definition**: The average grade a student has received across both the 1st and 2nd semesters.
   * **Calculation**:

avg\_grade=Curricular units 1st sem (grade)+Curricular units 2nd sem (grade)2

This gives an overall average of the grades across the two semesters.

1. **units\_without\_evaluation**:
   * **Definition**: This represents the total number of curricular units in which the student did not have any evaluations, combining both semesters.
   * **Calculation**:

units\_without\_evaluation=Curricular units 1st sem (without evaluations)+Curricular units 2nd sem (without evaluations)

1. **units\_with\_evaluation**:
   * **Definition**: This metric captures the total number of curricular units for which the student received evaluations, across both semesters.
   * **Calculation**:

units\_with\_evaluation=Curricular units 1st sem (evaluations)+Curricular units 2nd sem (evaluations)

1. **total\_credited**:
   * **Definition**: This is the total number of curricular units that a student has been credited for in both semesters.
   * **Calculation**:

total\_credited=Curricular units 1st sem (credited)+Curricular units 2nd sem (credited)

**Summary of Data Transformation and Parental Scores Calculation**

In this process, we performed several transformations to the dataset **academic4**, which involved:

1. **Assigning Weighted Scores to Mother's and Father's Occupations**:
   * Each occupation was assigned a weighted score based on its level of responsibility or education.
   * The scores ranged from 0 to 5, where:
     + **Higher-level occupations** like managers and professionals were given higher scores (e.g., 5 for managers, 4 for professionals).
     + **Lower-level occupations** such as unskilled labor and students were assigned lower scores (e.g., 1 for unskilled labor, 0 for students).
2. **Assigning Weighted Scores to Mother's and Father's Qualifications**:
   * Similar to occupations, qualifications were assigned scores based on their educational level:
     + **No education or low education** (e.g., no schooling, basic education) received a score of 0.
     + **Higher qualifications**, such as a **Doctorate** or **Master's Degree**, received higher scores (e.g., 5 for a Doctorate, 4 for a Master's).
     + **Bachelor’s degrees and technical courses** received middle-range scores (e.g., 3 for a Bachelor's degree, 2 for technical courses).
3. **Combining Parental Scores**:
   * A new variable, **parental\_scores**, was created by summing the weighted scores for both the **Mother's** and **Father's** occupations and qualifications.
   * This score represents a cumulative measure of parental education and occupation levels, giving insight into the socio-economic background of the student’s family.
4. **Normalization of Parental Scores**:
   * The **parental\_scores** were normalized to a scale of 1 to 5 using **Min-Max normalization**.
   * This transformation ensures that the parental scores fit within the desired range, where 1 represents the lowest possible parental score, and 5 represents the highest possible parental score. The formula for normalization used was:

Normalized Score=1+4×(Raw Score−Min)(Max−Min)\text{Normalized Score} = 1 + 4 \times \frac{(\text{Raw Score} - \text{Min})}{(\text{Max} - \text{Min})}Normalized Score=1+4×(Max−Min)(Raw Score−Min)​

* + - This ensures that all **parental\_scores** are mapped to a 1-5 scale, making it easier to interpret and analyze.

1. **Removing Irrelevant Columns**:
   * After assigning the weighted scores, the columns that were no longer needed, such as **Mother's occupation**, **Father's occupation**, **Mother's qualification**, **Father's qualification**, and others, were removed from the dataset. This was done to clean the dataset and focus only on relevant features for further analysis.