Category	Variable name	Data type	Missing data	Exploration and Unwanted observation
Demographic data	Marital status	Categorical	No	Keep
Demographic data	Nationality	Categorical	No	Drop. Highly corrected with variable International, and they are all imbalance.
Demographic data	Displaced	Categorical	No	Keep
Demographic data	Gender	Categorical	No	1. There are more female students than male students in the dataset. 2. The number of female students is 30% higher. 3. The number of dropout female and male students is almost equal. 4. The percentage of dropout male students is higher than the percentage of graduated male students
Demographic data	Age at enrollment	Numerical	No	1. The vast majority of students are 17-22 years old. 2. The number of students decreases as the students' age increases. 3. The highest graduation rate is the age group 17-22 with 72% of graduated students and 28% of dropped

				out students.
Demographic data	International	Categorical	No	1. The vast majority of students are Portuguese. 2. From the descriptive statistics we can see that one particular nationality appears 4,314. This is almost 98% of the whole dataset. Drop
Academic data at enrollment	Application mode	Categorical	No	Keep
Academic data at enrollment	Application order	Numerical	No	Drop
Academic data at enrollment	Course	Categorical	No	Drop. Removing columns with an absolute correlation coefficient less than 0.01.
Academic data at enrollment	Daytime/evening attendance	Categorical	No	Drop.
Academic data at enrollment	Previous qualification	Categorical	No	Drop. Highly imbalanced data.
Socioeconomic data	Mother's qualification	Categorical	No	Keep.
Socioeconomic data	Father's qualification	Categorical	No	Drop. Removing columns with an absolute correlation coefficient less than 0.01.
Socioeconomic data	Mother's occupation	Categorical	No	Keep
Socioeconomic data	Father's occupation	Categorical	No	Drop. Highly correlated with mother’s occupation.

Socioeconomic data	Educational special needs	Categorical	No	Drop. Removing columns with an absolute correlation coefficient less than 0.01.
Socioeconomic data	Debtor	Categorical	No	Keep
Socioeconomic data	Tuition fees up to date	Categorical	No	Keep
Socioeconomic data	Scholarship holder	Categorical	No	Keep
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numerical	No	Drop. Does not contribute a lot of differences in dropout and graduate
Academic data at the end of 1st semester	Curricular units 1st sem (enrolled)	Numerical	No	Keep. Combine with Curricular units 2nd sem (enrolled)
Academic data at the end of 1st semester	Curricular units 1st sem (evaluations)	Numerical	No	Drop. Does not contribute a lot of differences in dropout and graduate
Academic data at the end of 1st semester	Curricular units 1st sem (approved)	Numerical	No	Keep. Combine with Curricular units 2nd sem (approved)
Academic data at the end of 1st semester	Curricular units 1st sem (grade)	Numerical	No	Keep. Combine with Curricular units 2nd sem (grade)
Academic data at the end of 1st semester	Curricular units 1st sem (without evaluations)	Numerical	No	Drop. Does not contribute a lot of differences in dropout and graduate
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numerical	No	Drop. Does not contribute a lot of differences in dropout and graduate
Academic data at the end of 2nd semester	Curricular units 2nd sem (enrolled)	Numerical	No	Keep. Combine with Curricular units 2nd sem (enrolled)

Academic data at the end of 2nd semester	Curricular units 2nd sem (evaluations)	Numerical	No	Drop. Does not contribute a lot of differences in dropout and graduate.
Academic data at the end of 2nd semester	Curricular units 2nd sem (approved)	Numerical	No	Keep. Combine with Curricular units 1st sem (approved)
Academic data at the end of 2nd semester	Curricular units 2nd sem (grade)	Numerical	No	Keep. Combine with Curricular units 1st sem (grade)
Academic data at the end of 2nd semester	Curricular units 2nd sem (without evaluations)	Numerical	No	Drop. Does not contribute a lot of differences in dropout and graduate
Macroeconomic data	Unemployment rate	Numerical	No	Drop. Removing columns with an absolute correlation coefficient less than 0.01.
Macroeconomic data	Inflation rate	Numerical	No	Keep
Macroeconomic data	GDP	Numerical	No	Keep
	Target	Categorical	No	Graduate, Dropout, Enrollment. I will drop the enrollment students in the EDA The total number of graduated and dropout students is 2209 + 1421 = 3630.