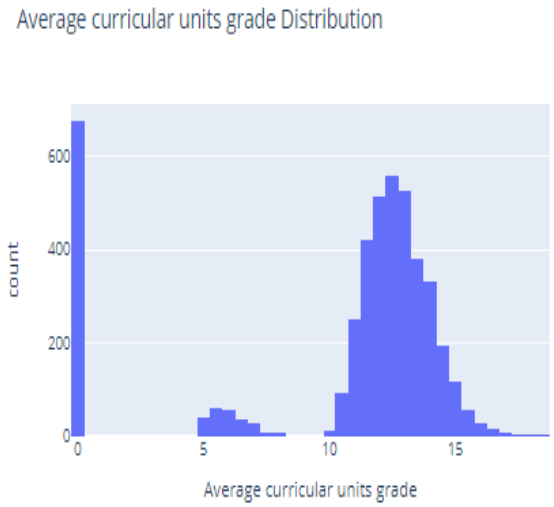
DATA EXPLORATION ANALYSIS

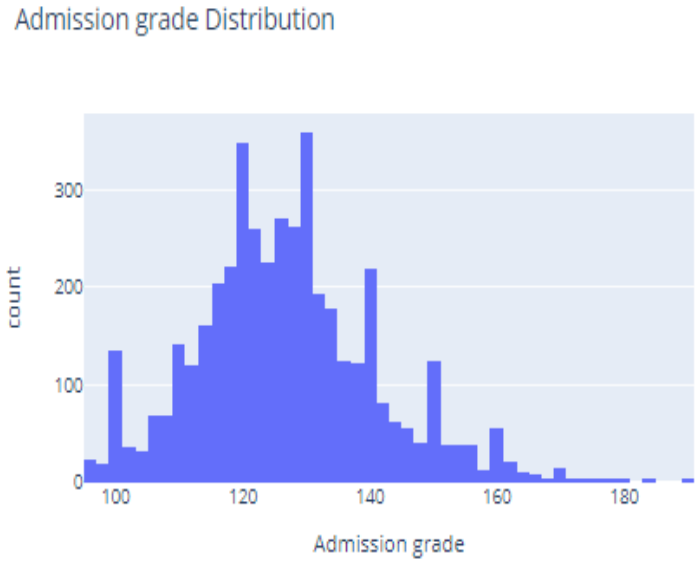
How the Data Was Collected

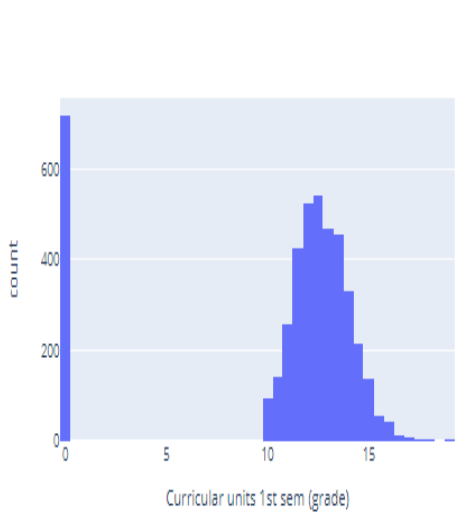
The dataset was collected from 3SignetDatasets. This dataset contains summary students' dropout rates for 4402 students, as of 20th of August 2024. It was downloaded into a desktop folder which was later loaded into pandas dataframe for analysis.

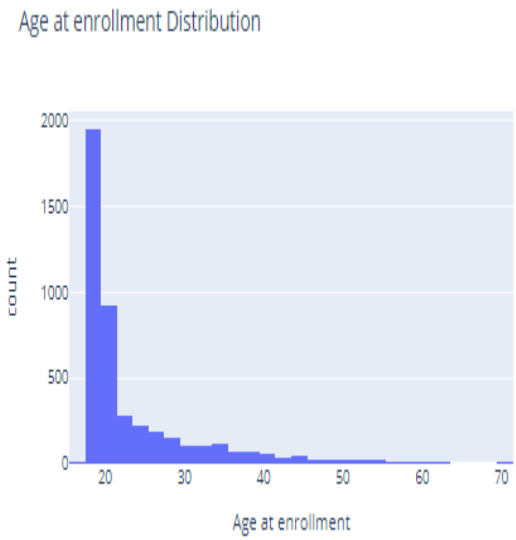
The features identified for the analysis are Age at enrolment, Average curricular units grade, Unemployment rate, Inflation rate, GDP ,Previous qualification (grade),Admission grade, Curricular units 1st sem (grade),Curricular units 2nd sem (grade), Daytime/evening attendance, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, International, Marital status, Target, Previous qualification, Nationality, Course, Curricular units 1st sem (credited),Curricular units 1st sem (enrolled),Curricular units 1st sem (evaluations),Curricular units 1st sem (approved),Curricular units 1st sem (without evaluations),Curricular units 2nd sem (credited),Curricular units 2nd sem (enrolled),Curricular units 2nd sem (evaluations),Curricular units 2nd sem (approved),Curricular units 2nd sem (without evaluations).The reason for choosing these features is simply because we want to identify the risks associated with students dropout. The key findings are summarized below:

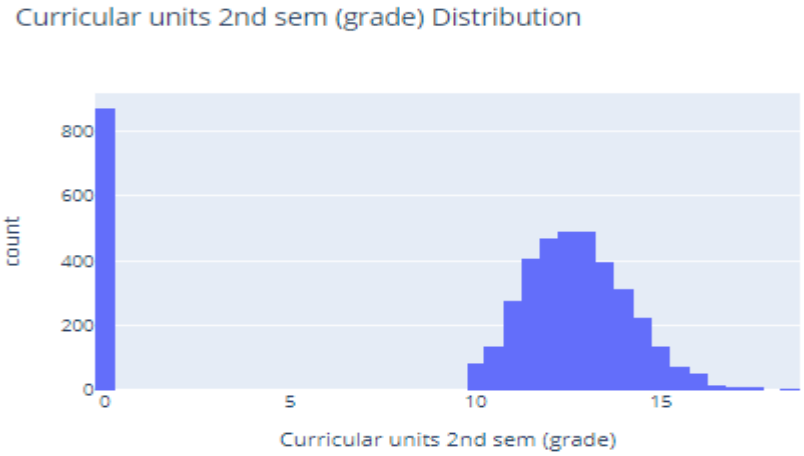
Screenshots of Pandas-Profiling Reports
Histogram showing Distributions of Numerical Variable





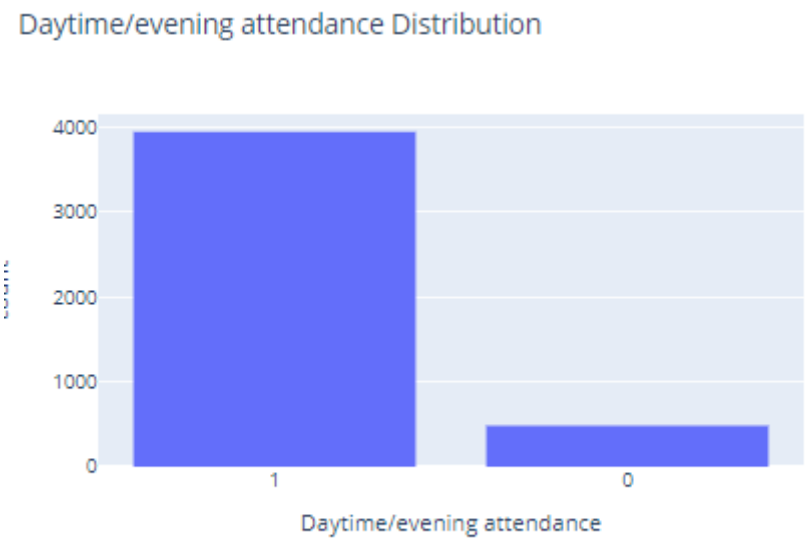




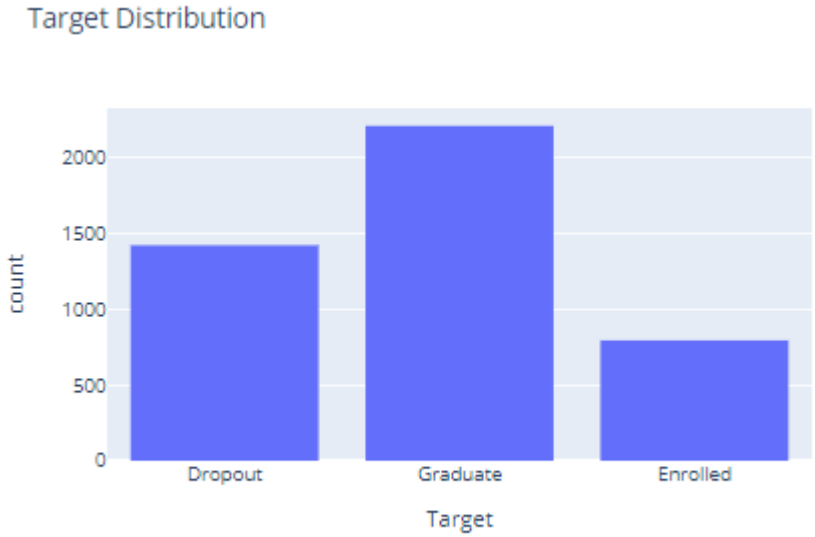


The grade distributions show that students follow the same pattern in their academic grades. The data is mostly centered and follow a normal distribution. The age column show that the students are mostly less than 20 years old.

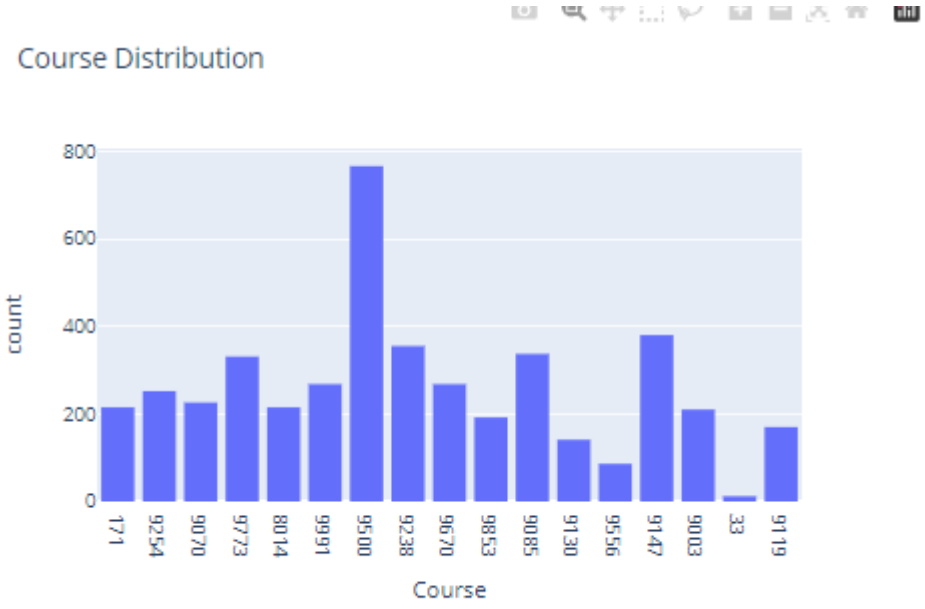
The box plot of the first and second show the maintain the same grades.



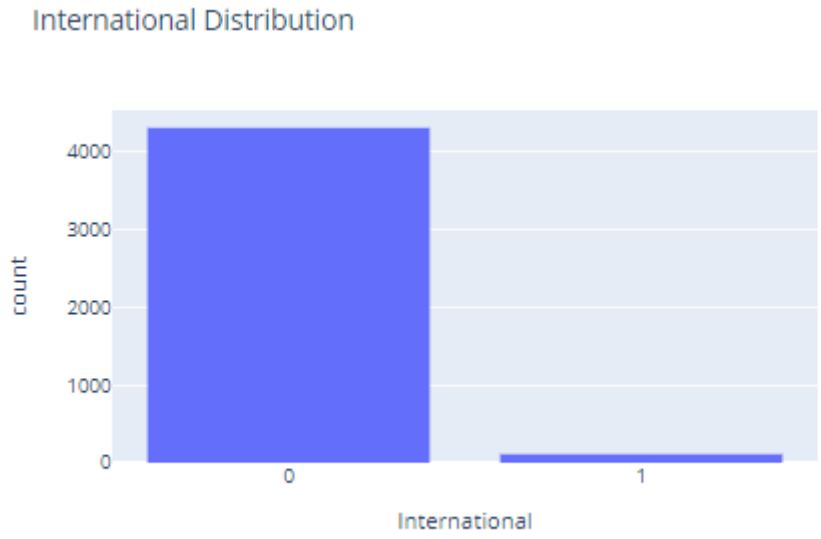
This shows that people who attend evening classes(0) are 4 times more than those who attend daytime classes



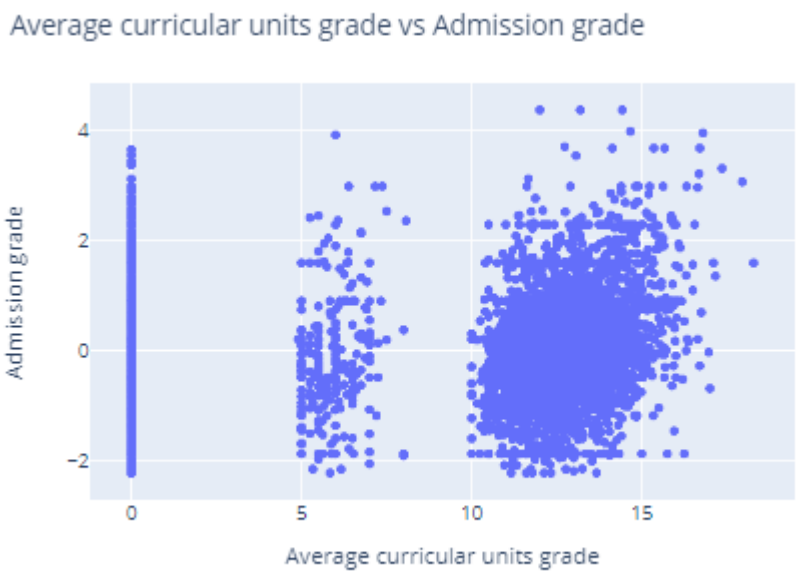
Students in the target columns are more of Graduates but the dropouts are more than those enrolled.



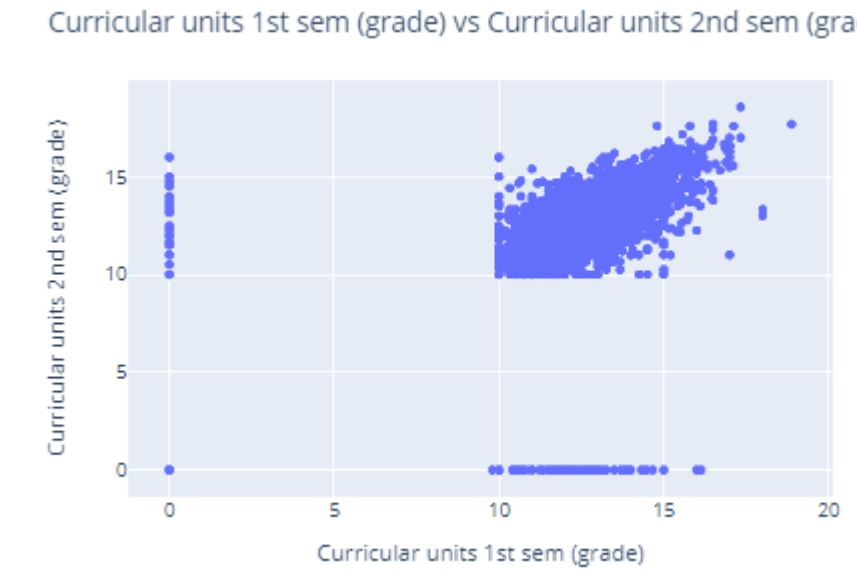
Course 9500 has the highest population of students with 33 being the least.



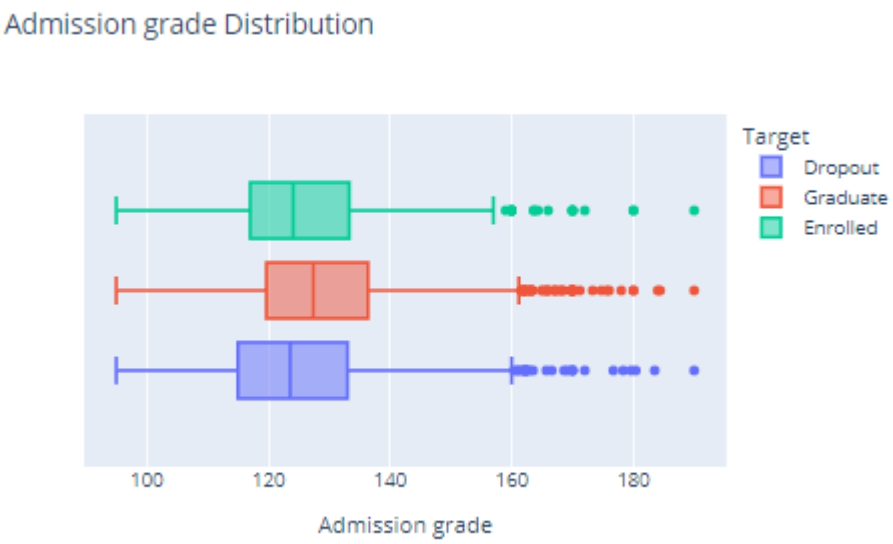
Non international students are 4x much more than international students



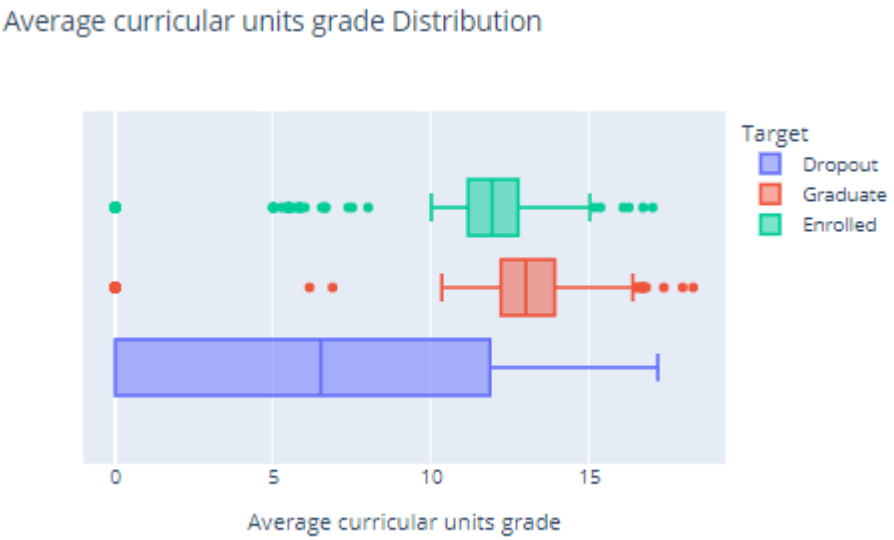
Average grades and Admission grades are correlated. The above plot shows that as admission grades increases, the average grades increases.



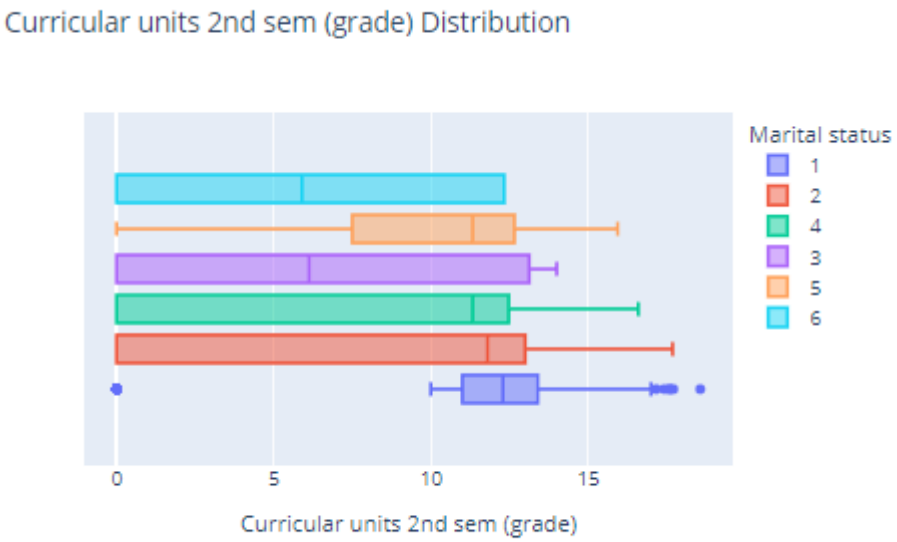
Curricular units 1st sem (grade) and Curricular units 2nd sem (grade) are also correlated.



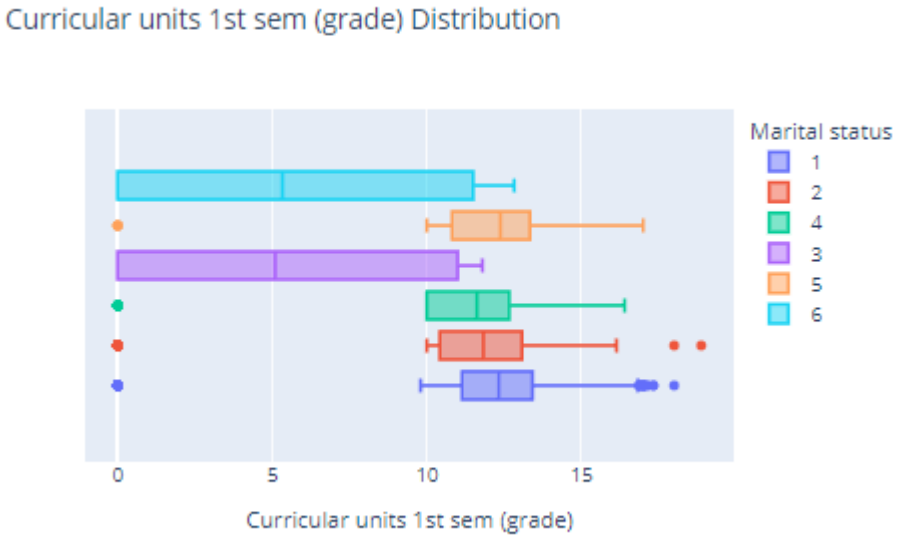
The distribution shows that Graduate students have the highest median but the distributions of grades are equal for all category.



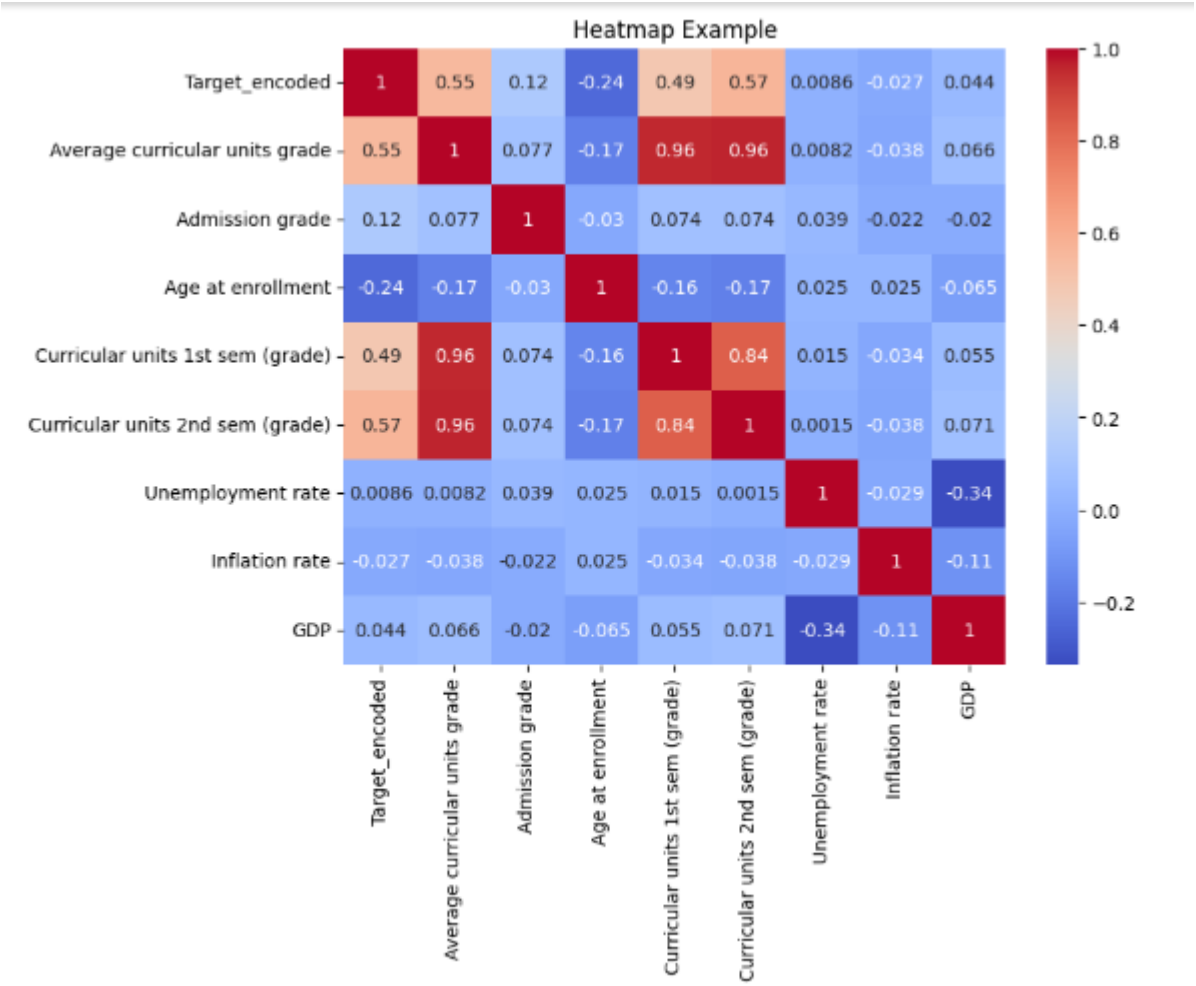
The distribution shows that dropouts perform lower than other categories on average.



The distribution shows that students belonging to group 1 in the marital status performed higher than other categories in the 2nd semester grade



The distribution shows that students belonging to group 1 in the marital status also performed higher than other categories in the 1st semester grade



This shows the correlation between numerical variables such as the grades and shows that the 1st semester grades and 2nd semester grades are highly correlated and with their average.

Hypothesis Test

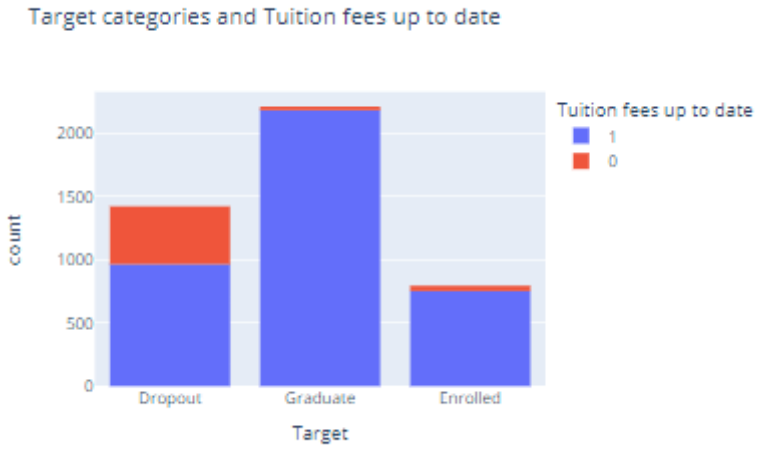
The following hypothesis tests were carried out to test if there are any association the categories with Ho meaning the null hypothesis and Ha meaning the alternative hypothesis.

1. Target categories and Tuition fees up

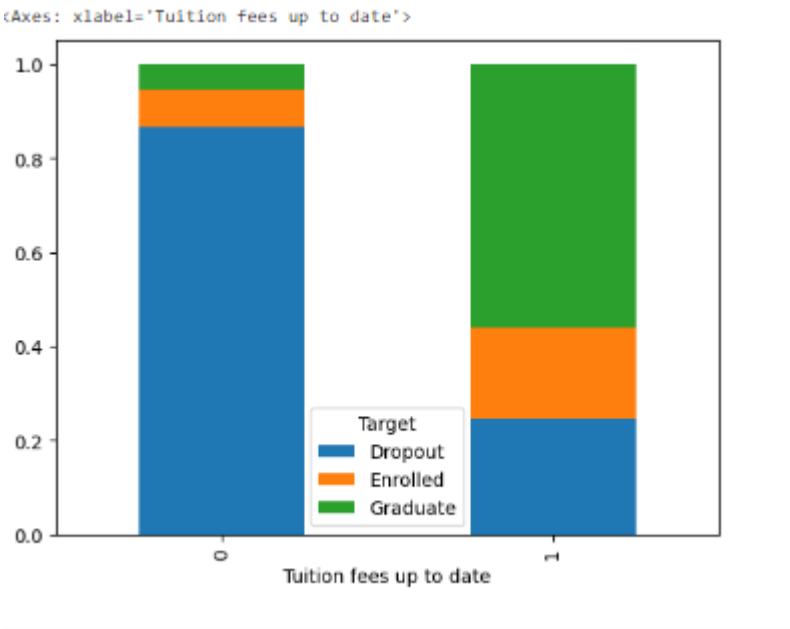
Ho: Target categories and Tuition fees up to date are independent
Ha: Target categories and Tuition fees up to date are not independent

Target	proportion
Graduate	0.499322
Dropout	0.321203
Enrolled	0.179476

This shows the proportions of each category in the target column



The above shows the distributions



The stacked bar plots shows they are not independent since the splits are not on the same level

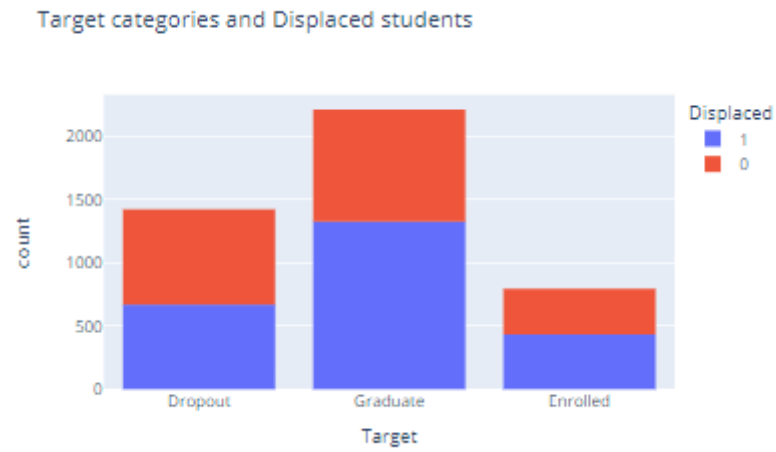
	test	lambda	chi2	dof	pval	cramer	power
0	pearson	1.0	823.552724	2.0	1.471628e-179	0.431458	1.0

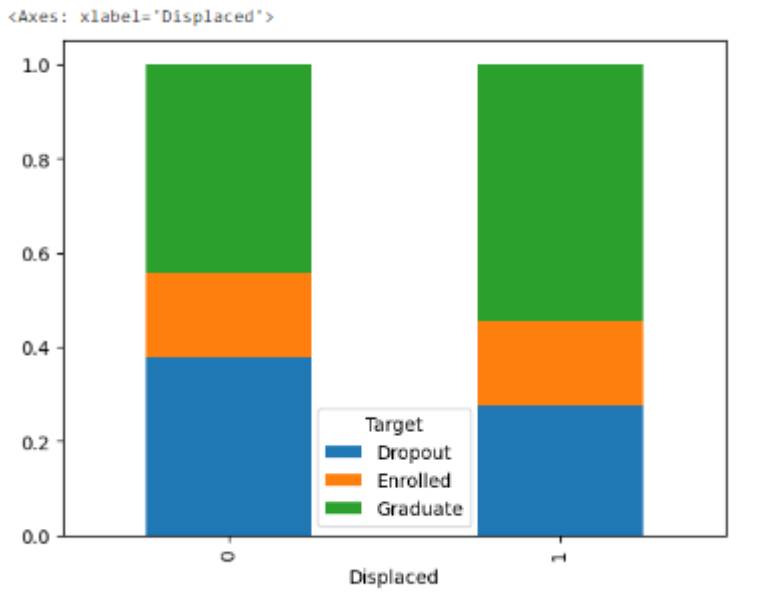
The p-value is less than a significant level of 0.05 so we reject the null hypothesis and conclude that the target categories and the Tuition fees up to date are not independent

Pearson correlation was used for the test and a significance level of 0.05 and since it was less than 0.05, we reject the hypothesis and conclude that they are not independent.

2. Target categories and Displaced

Ho: Target categories and Displaced students are independent
Ha: Target categories and Displaced students are not independent





```
      test  lambda       chi2  dof          pval    cramer  power
0  pearson     1.0  57.754195  2.0  2.876311e-13  0.114257    1.0
```

The p-value is less than a significant level of 0.05 so we reject the null hypothesis and conclude that the target categories and the Displaced students are not independent

The test between Target categories and Displaced showed the same result as the test between Target categories and Tuition fees up to date. So we conclude that they are not independent.

Principal Component Analysis

This was also carried out to find the principal components in the numerical features. And below shows their respective results

	Average curricular units grade_encoded	Admission grade_encoded	Age at enrollment_encoded	Curricular units 1st sem (grade)_encoded	Curricular units 2nd sem (grade)_encoded	Unemployment rate_encoded	Inflation rate_encoded	GDP_encoded
0	-0.583431	-0.069453	1.538666e-01	-0.557784	-0.560536	4.161279e-03	3.937533e-02	-6.494787e-02
1	0.039114	0.116986	1.229428e-01	0.047473	0.028211	6.742379e-01	1.567142e-01	-6.981860e-01
2	0.060312	-0.513089	1.837693e-01	0.060645	0.055172	-2.299343e-01	7.950392e-01	-8.747390e-02
3	0.046353	-0.780910	2.727351e-01	0.047737	0.041354	1.929558e-01	-5.219080e-01	-6.117378e-03
4	0.066777	0.329222	9.226797e-01	0.071611	0.056934	-1.299961e-01	-4.259859e-02	9.344911e-02
5	0.015808	0.004654	-3.919032e-02	0.016075	0.014294	-6.620697e-01	-2.599917e-01	-7.012830e-01
6	0.017997	-0.000630	1.344753e-02	-0.715383	0.698262	6.992267e-03	2.163631e-03	-9.919187e-03
7	-0.804524	-0.000000	5.551115e-17	0.404360	0.435010	-4.770490e-18	5.204170e-18	-4.163336e-17