**Raport**

**Author:** 280462 Bartosz Wacławiak

**Date:** xxx

**Course:** MSiD - laboratories

**Dataset:** [**Predict Student’s Dropout and Academic Success**](https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success)

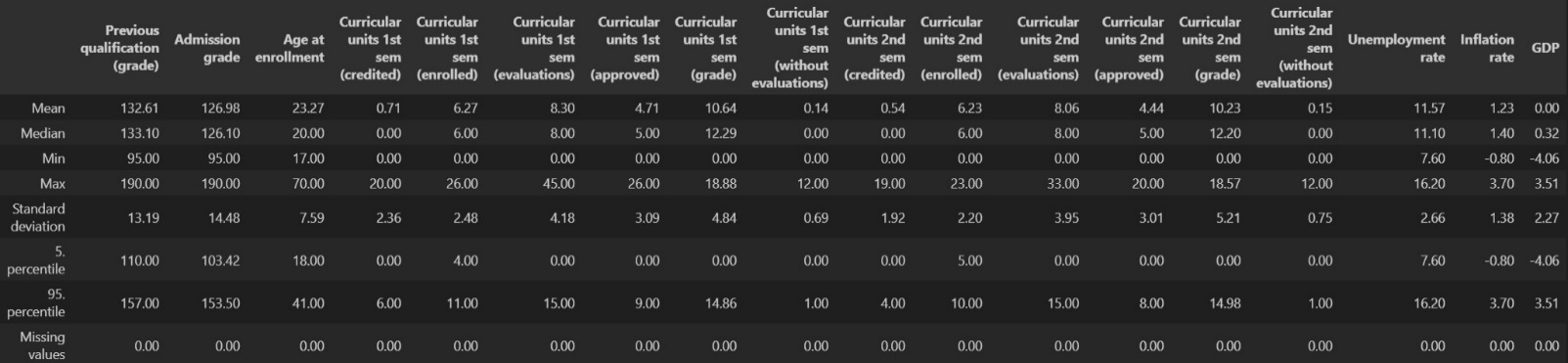
**Part 1**

This section presents a comprehensive visual analysis of student academic behavior and performance using a variety of statistical charts. The goal is to uncover meaningful patterns and relationships between student characteristics, academic outcomes, and external factors such as economic indicators.

The charts explore variables such as age at enrollment, admission grades, course performance, financial status (debtor or scholarship holder), and macroeconomic context (GDP, unemployment, inflation). By analyzing these factors across different student groups (dropouts, graduates, and currently enrolled), we aim to identify trends that may explain academic success or failure and provide insights for better educational planning and policy-making.

Each visual is accompanied by a concise set of observations that highlight key findings and support data-driven interpretation.

**Feature statistics**

**Numeric features analysis**

**Most common categories**

* Application mode -> 1st phase - general contingent (38.61%)
* Application order -> 2nd choice (68.40%)
* Course -> Nursing (17.31%)
* Daytime evening attendance -> Daytime (89.08%)
* Debtor -> No (88.63%)
* Displaced -> Yes (54.84%)
* Educational special needs -> No (98.85%)
* Father's occupation -> Unskilled Workers (22.83%)
* Father's qualification -> Basic Ed 1st Cycle (4th/5th) (27.33%)
* Gender -> Female (64.83%)
* International -> No (97.51%)
* Matrimonial status -> single (88.58%)
* Mother's occupation -> Unskilled Workers (35.65%)
* Mother's qualification -> Secondary Education (24.16%)
* Nationality -> Portuguese (97.51%)
* Previous qualification -> Secondary education (84.02%)
* Scholarship holder -> No (75.16%)
* Target -> Graduate (49.93%)
* Tuition fees up to date -> Yes (88.07%)

**Chart types**

**Heatmap**

A heatmap is a graphical representation of data where individual values are represented by colors. In the context of correlation analysis, a heatmap visually displays the strength and direction of relationships between multiple variables. Darker or more intense colors indicate stronger correlations (positive or negative), making it easy to identify patterns and dependencies in the dataset.

**PCA**

PCA (Principal Component Analysis) is a dimensionality reduction technique used to simplify large datasets while preserving as much variance as possible. It transforms the original features into a new set of uncorrelated variables called principal components. These components help visualize complex data in two or three dimensions and are useful for detecting structure, clusters, or outliers.

**Histogram**

A histogram is a type of bar chart that shows the distribution of a numerical variable by dividing the data into intervals (called bins) and counting how many values fall into each interval. It provides a visual overview of the shape, spread, and central tendency of the data, helping to identify patterns.

**Error plot**

An error plot is a visualization that shows the variability or uncertainty around data points, often using vertical or horizontal bars to represent error margins. These may reflect standard deviation, confidence intervals, or standard error, helping to interpret the reliability and precision of the displayed values.

**Box plot**

A box plot (also known as a box-and-whisker plot) is used to summarize the distribution of a dataset based on five key statistics: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It highlights central tendency, variability, and potential outliers, making it easy to compare distributions across different groups.

**Violin plot**

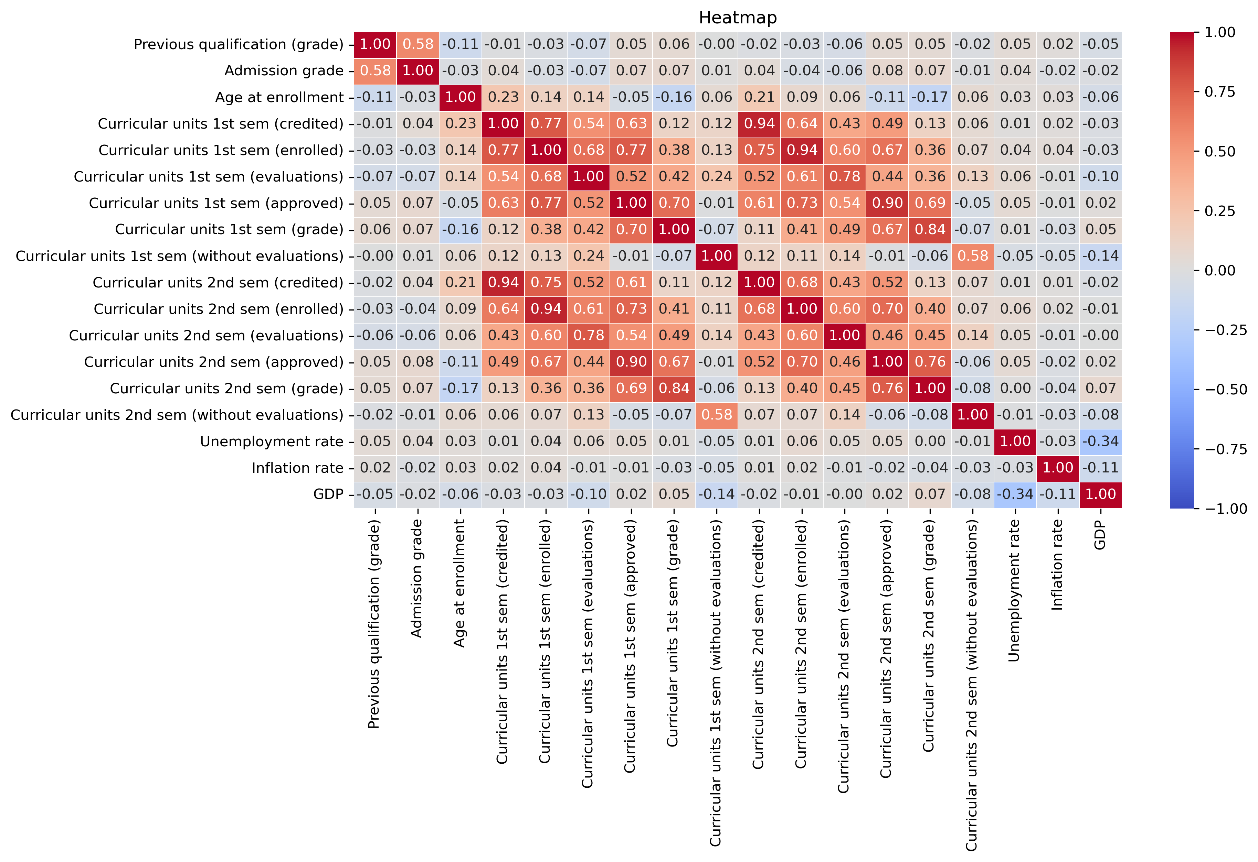
A violin plot is a combination of a box plot and a kernel density plot. It displays the distribution, probability density, and summary statistics of a dataset. The width of the violin shape shows the frequency of data points at different values, while the internal box plot provides information on median and quartiles. Violin plots are useful for comparing multiple groups and visualizing the shape of the data distribution.

**Regression plot**

A regression plot is a visualization that shows the relationship between two variables along with a fitted regression line, which models the trend in the data. It often includes a confidence interval around the line to indicate the uncertainty of the estimate. Regression plots are useful for identifying linear or nonlinear trends, correlations, and the strength of associations between variables.

**Data overview**

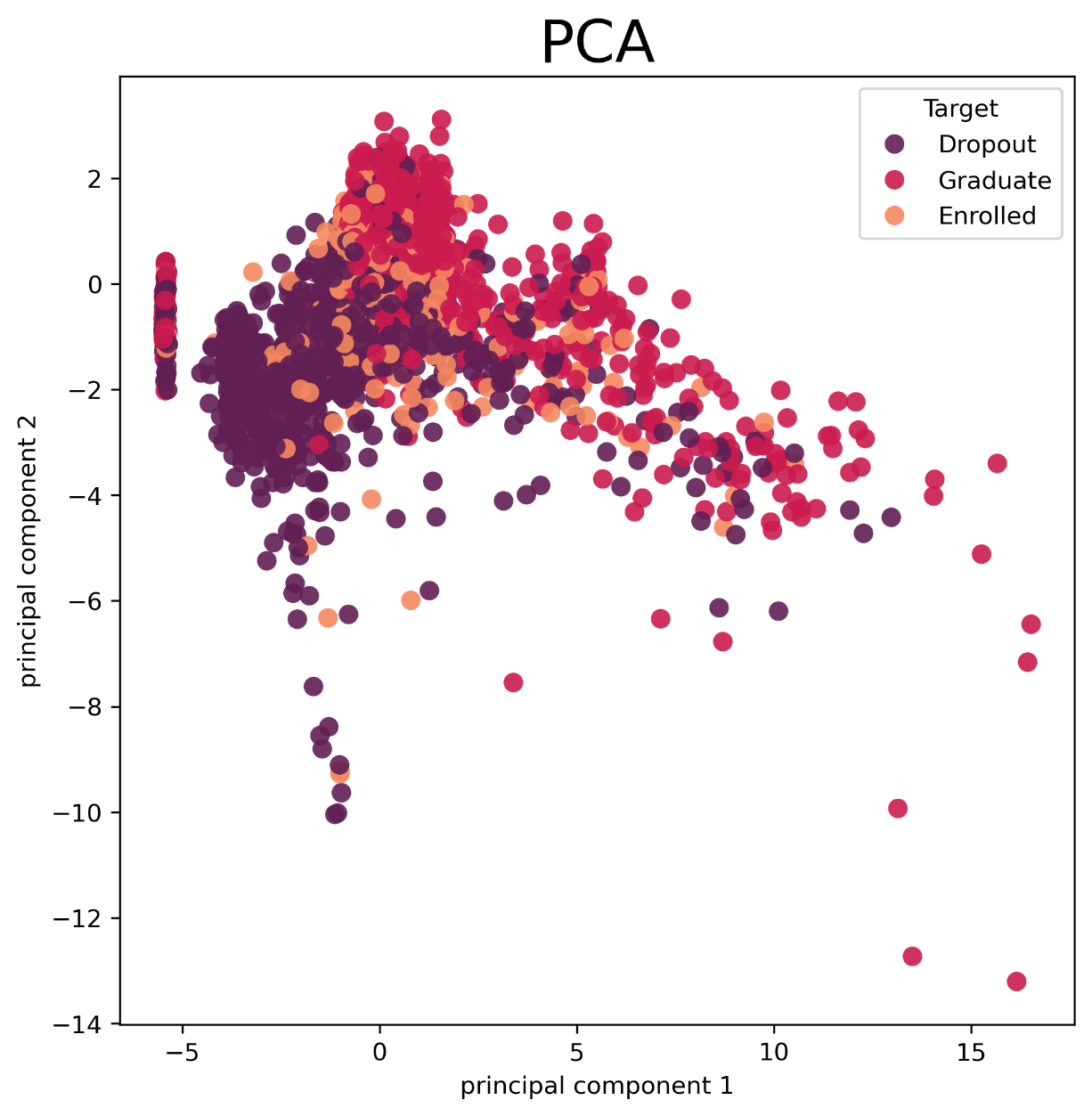
**Heatmap**

****

Observations

* The parameters related to curricular units in the 1st semester and in the 2nd semester (credited, enrolled, evaluations, approved, grade) show strong mutual correlation, which is expected since they describe related aspects of academic performance.
* The is a strong positive correlation between admission grade and previous qualification grade (0.58). This suggests consistency in academic performance before and during university admission.
* A slight negative correlation is observed between age at enrollment and academic performance indicators, particularly with grades. This may indicate that younger students tend to perform slightly better.
* A clear negative correlation between GDP and unemployment rate is observed (−0.75), reflecting a typical economic relationship where higher GDP often aligns with lower unemployment.
* The **inflation rate does not show significant correlation** with either GDP or unemployment rate, which is somewhat **unexpected** and may require further investigation.

**PCA**

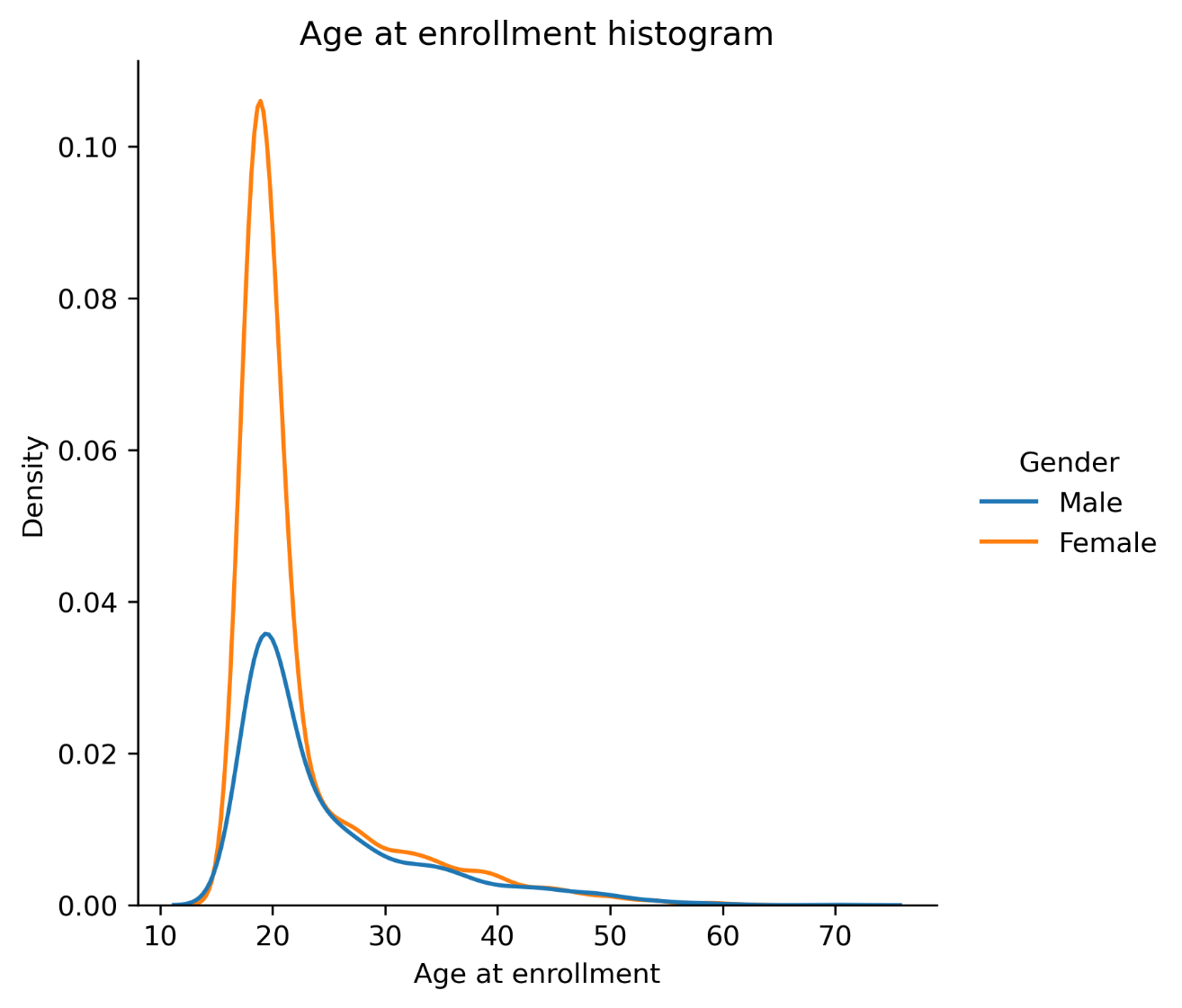
****

Observations

* The dropout students category is primarily located on the left side of the chart, indicating a clustering pattern along lower values of the first principal component.
* The graduate students category tends to be distributed toward the upper regions of the chart, showing some separation from other categories in terms of both principal components.
* The enrolled students category is more dispersed across the plot but is often positioned between the dropout and graduate student clusters, reflecting transitional characteristics.

**Correlations**

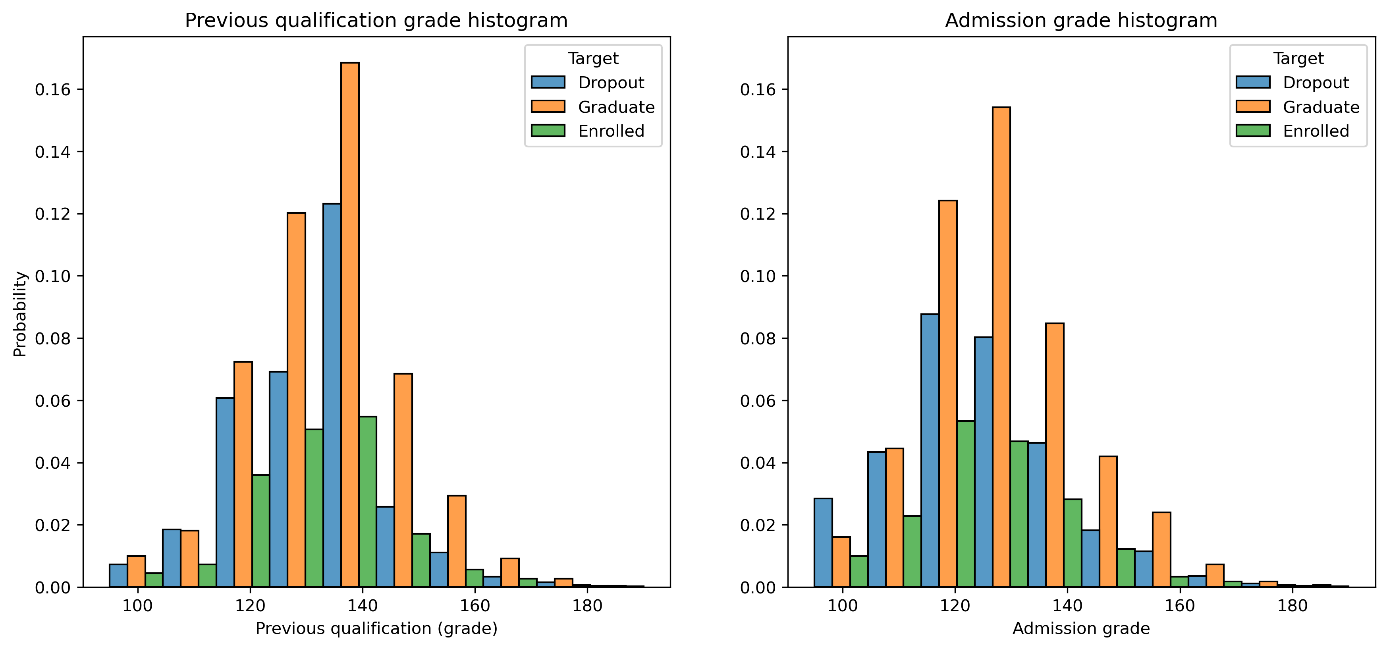
**Gender**

****

Observations

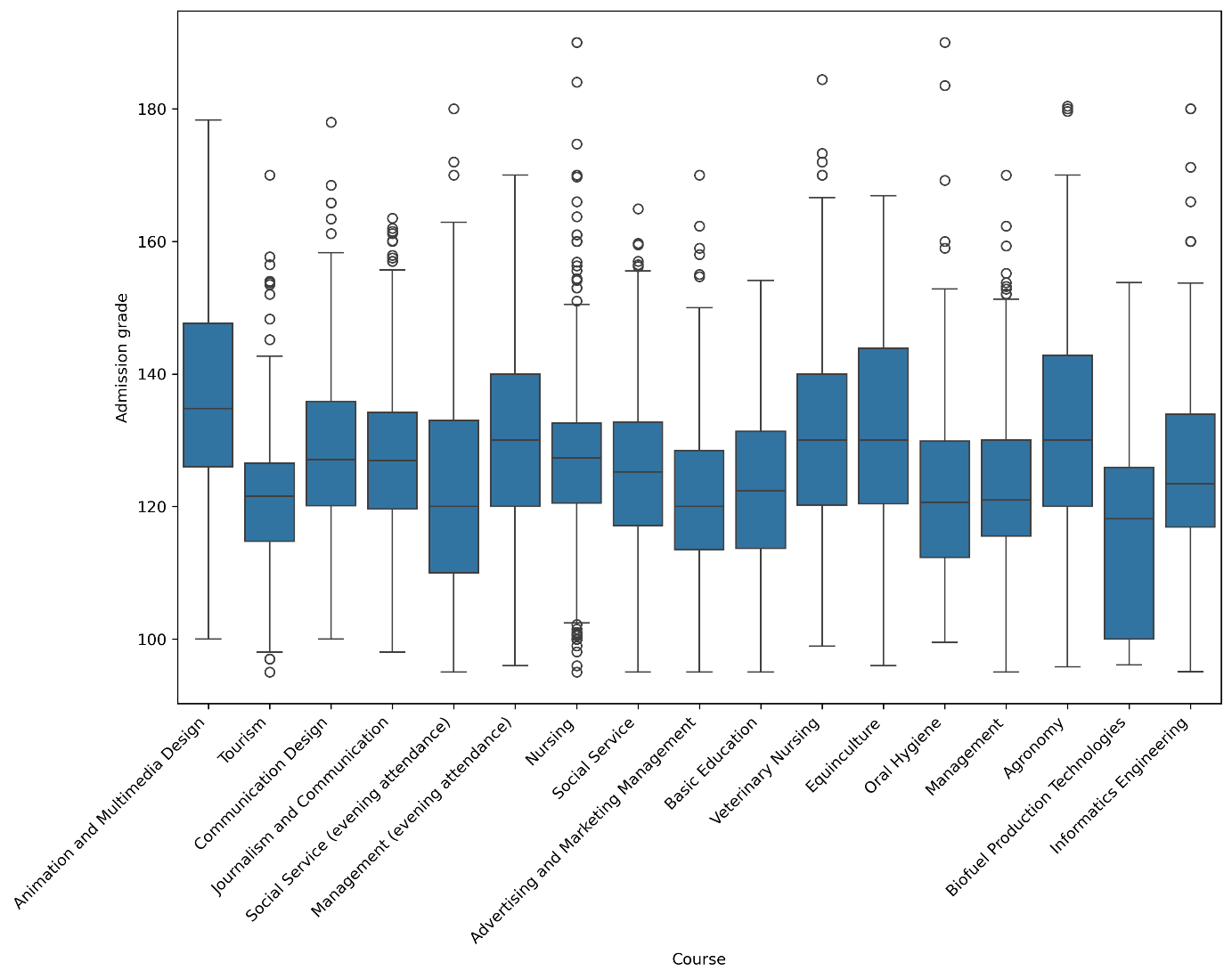
* The majority of students, regardless of gender, enroll around the age of 18 to 20, as shown by the sharp peak in that range.
* The distribution for female students is higher in density than for male students, suggesting there are more female students in the dataset.
* Both male and female age distributions are right-skewed, meaning there are fewer older students, but a small number do enroll later in life—even beyond age 50.
* After the main peak, the number of enrollments steadily declines for both genders, though the distributions remain somewhat similar in shape.

**Previous qualification and admission grades**

****

Observations

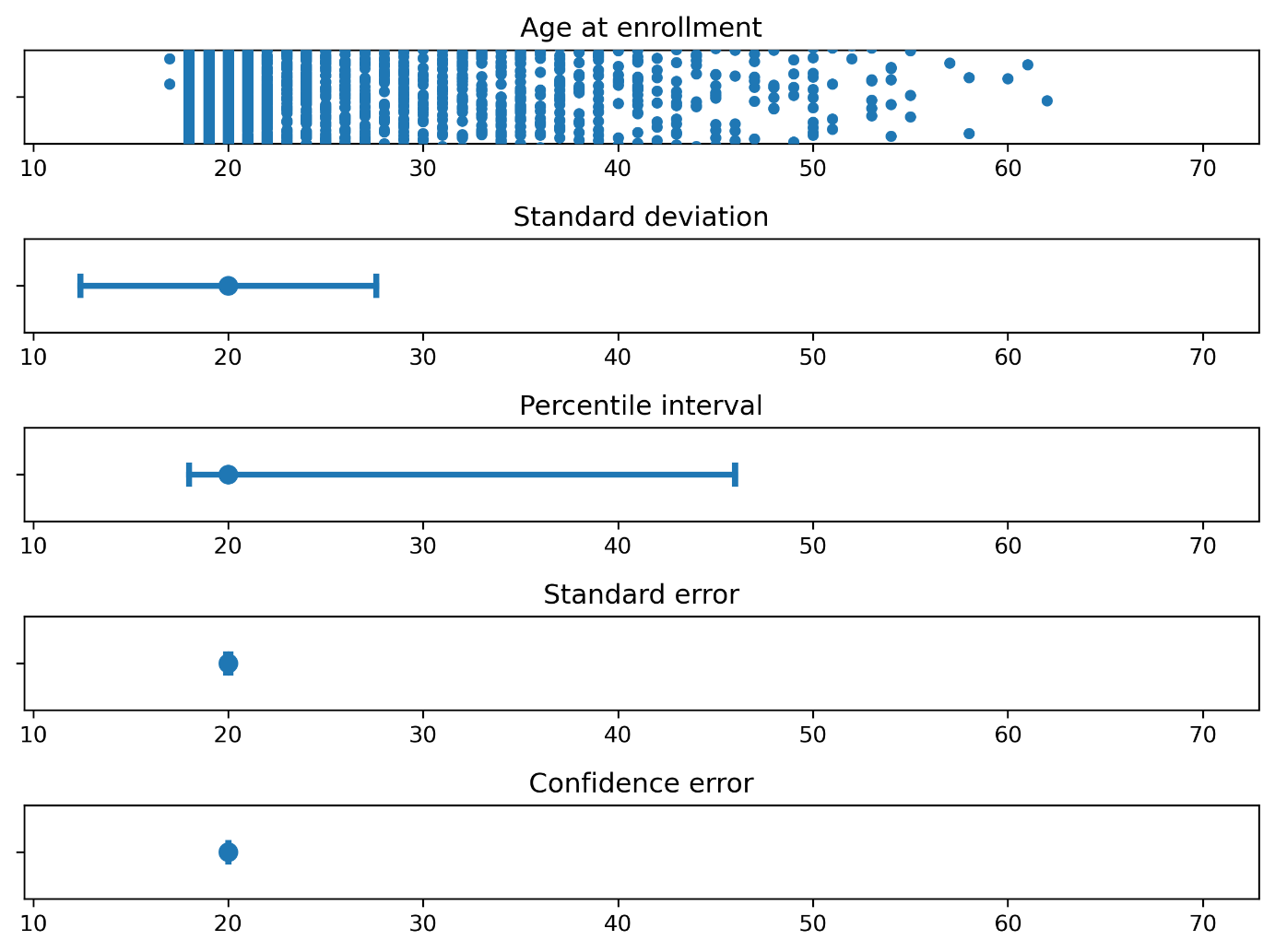
* The distribution of both previous qualification grades and admission grades follows a bell-shaped curve, with most students scoring between 120 and 140.
* Dropout students (blue bars) are more present in the lower grade ranges, particularly around 110–130, which may suggest a link between lower grades and higher dropout rates.
* Overall, students with higher previous and admission grades are more likely to graduate, indicating a potential predictive relationship between early academic performance and university outcomes.



Observations

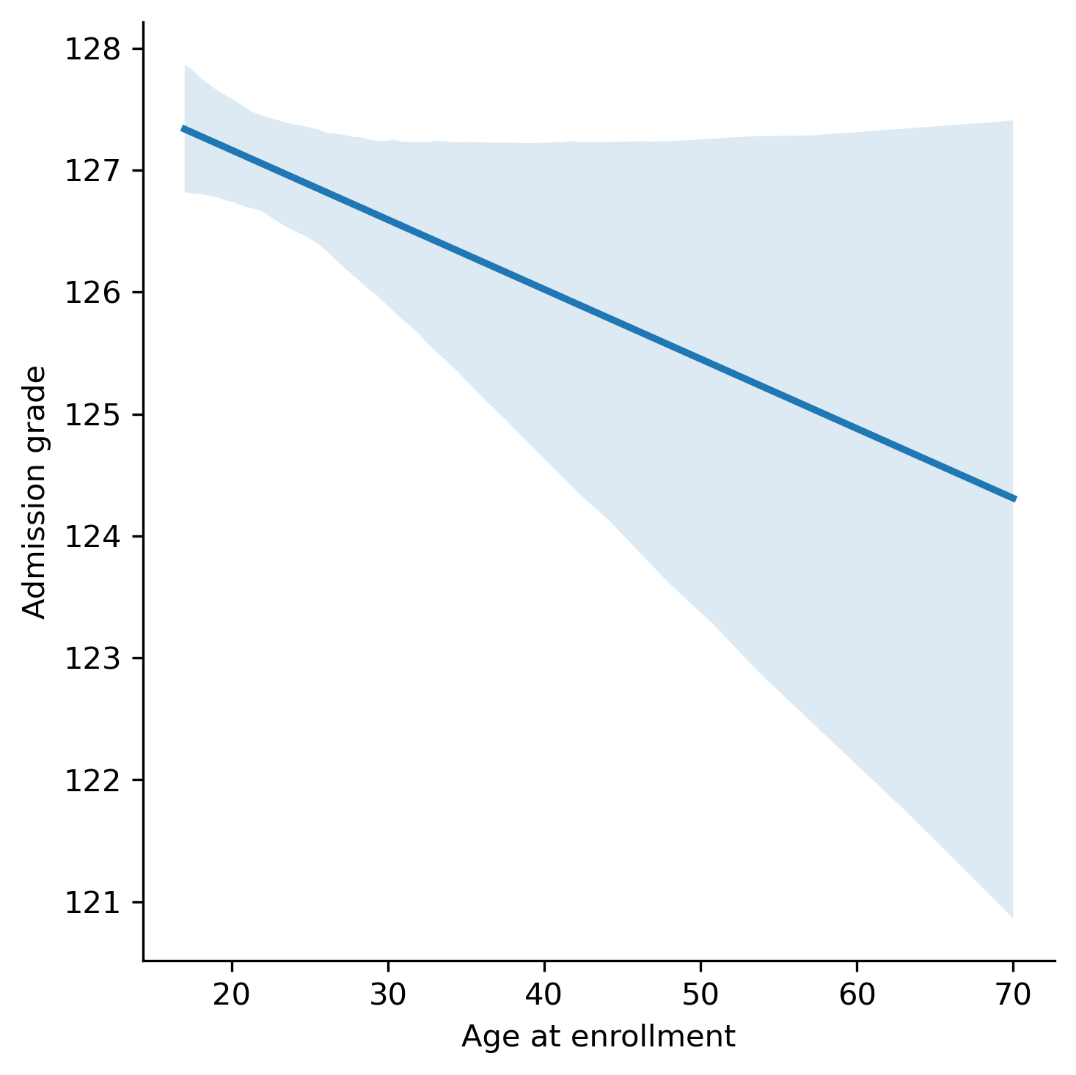
* Students enrolled in the Animation and Multimedia Design course have the highest admission grades on average, with relatively few low-performing outliers.
* In contrast, students in Biofuel Production Technologies tend to have lower admission grades, with a more compressed distribution and a lower median compared to other courses.
* Several courses, such as Tourism, Journalism and Communication, and Management, show wider variability in admission grades, indicating more diverse student performance at entry.
* Courses like Oral Hygiene and Equiniculture display relatively high median scores and narrower interquartile ranges, suggesting more consistent performance among admitted students.
* The presence of outliers across many courses indicates that while most students fall within a typical range, there are exceptional cases of both high and low performance across all programs.

**Age at enrollement**



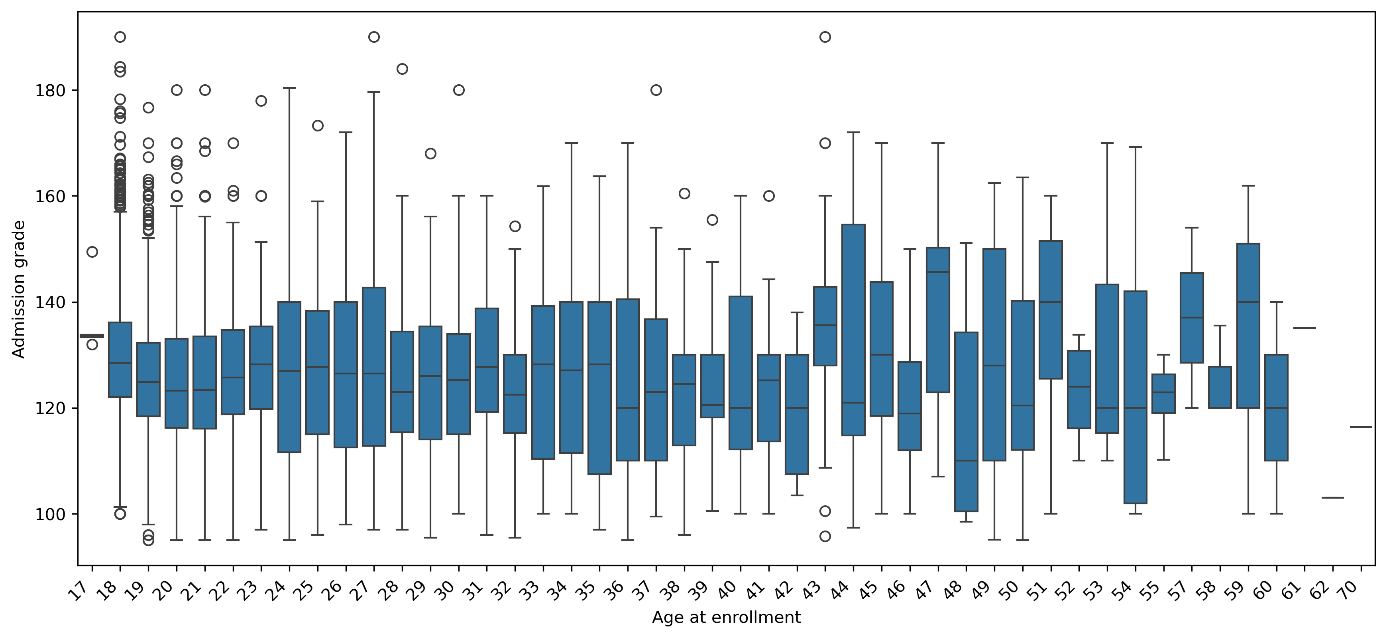
Obsevations

* Most students enroll between the ages of 18 and 25, as shown in the dot plot and supported by all statistical intervals (standard deviation, percentile, and error bounds).
* The standard deviation and percentile interval indicate that the bulk of enrollment ages fall roughly between 18 and 45, with some outliers extending to age 70.



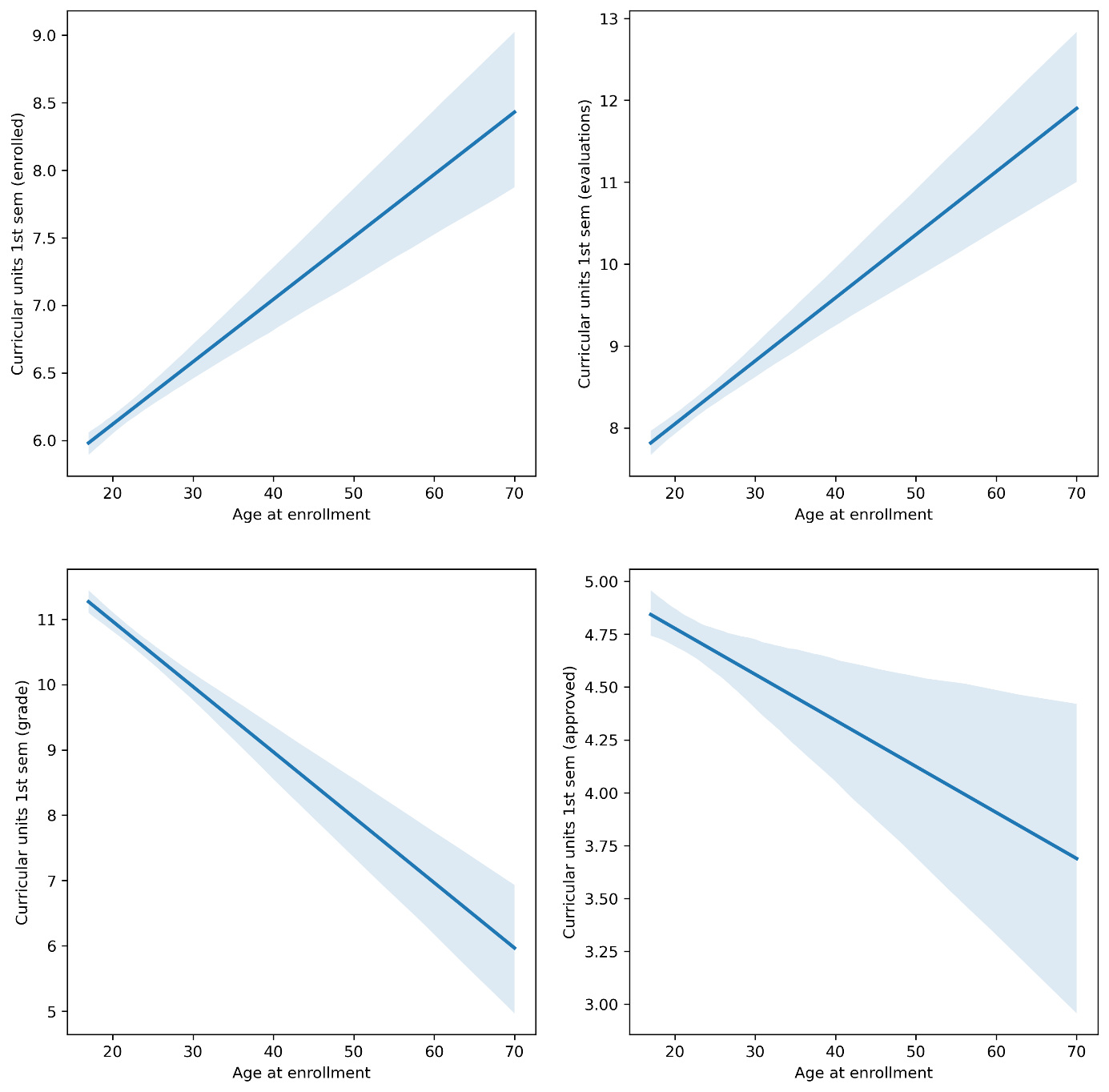
Observations

* The regression plot reveals a slight negative correlation between age at enrollment and admission grade — suggesting that students who enroll at an older age tend to have slightly lower admission grades on average.
* However, the confidence interval around the regression line widens significantly for older students, indicating less certainty in the trend due to sparser data in those age ranges.



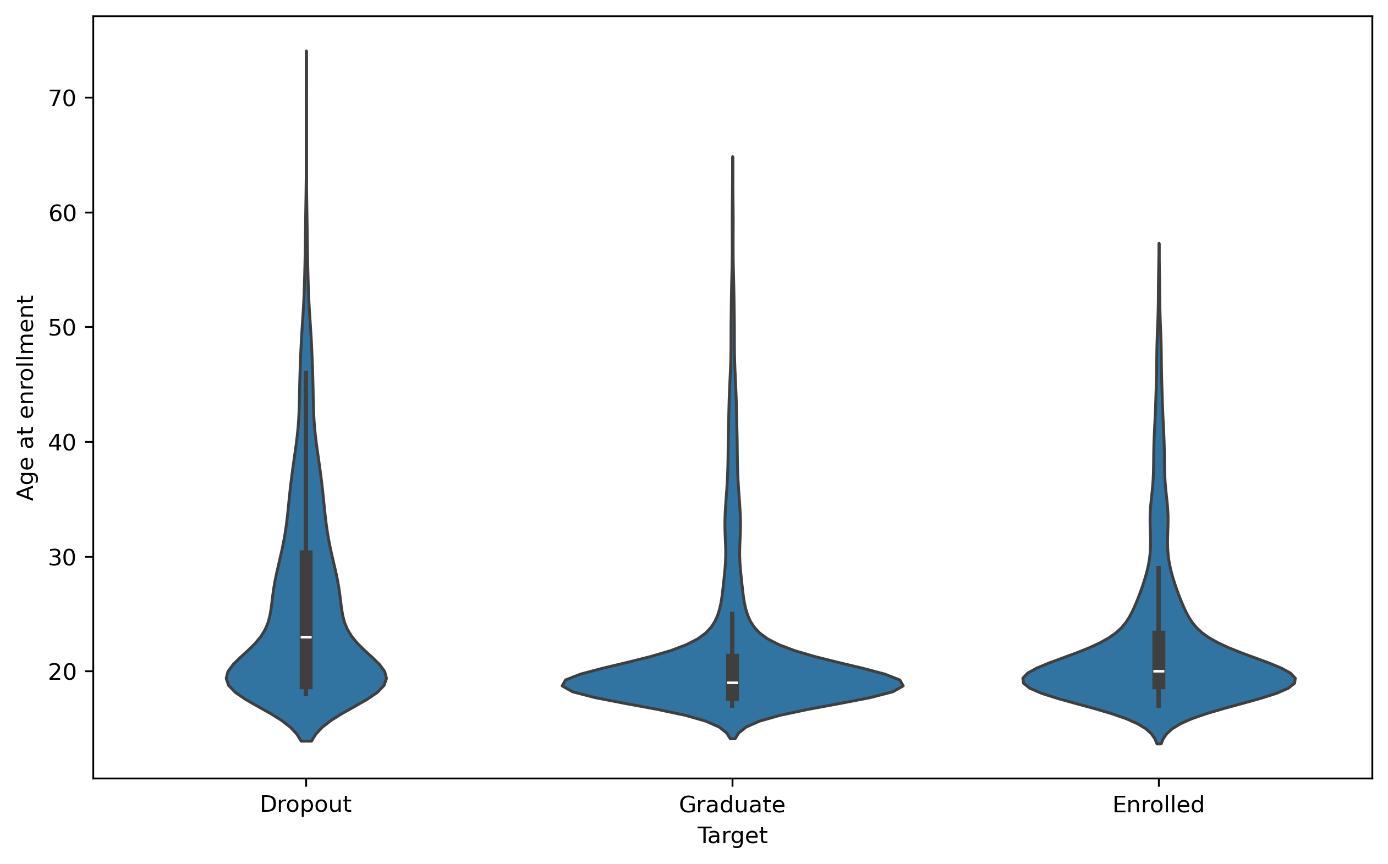
Observations

* The boxplot of admission grade by age supports this trend: younger students (especially those around age 18–21) tend to have higher median grades, while older age groups show more variability and often lower medians.
* Overall, while the trend suggests older students perform slightly worse in admission scores, the effect may be partially due to the smaller sample size for non-traditional-age students.



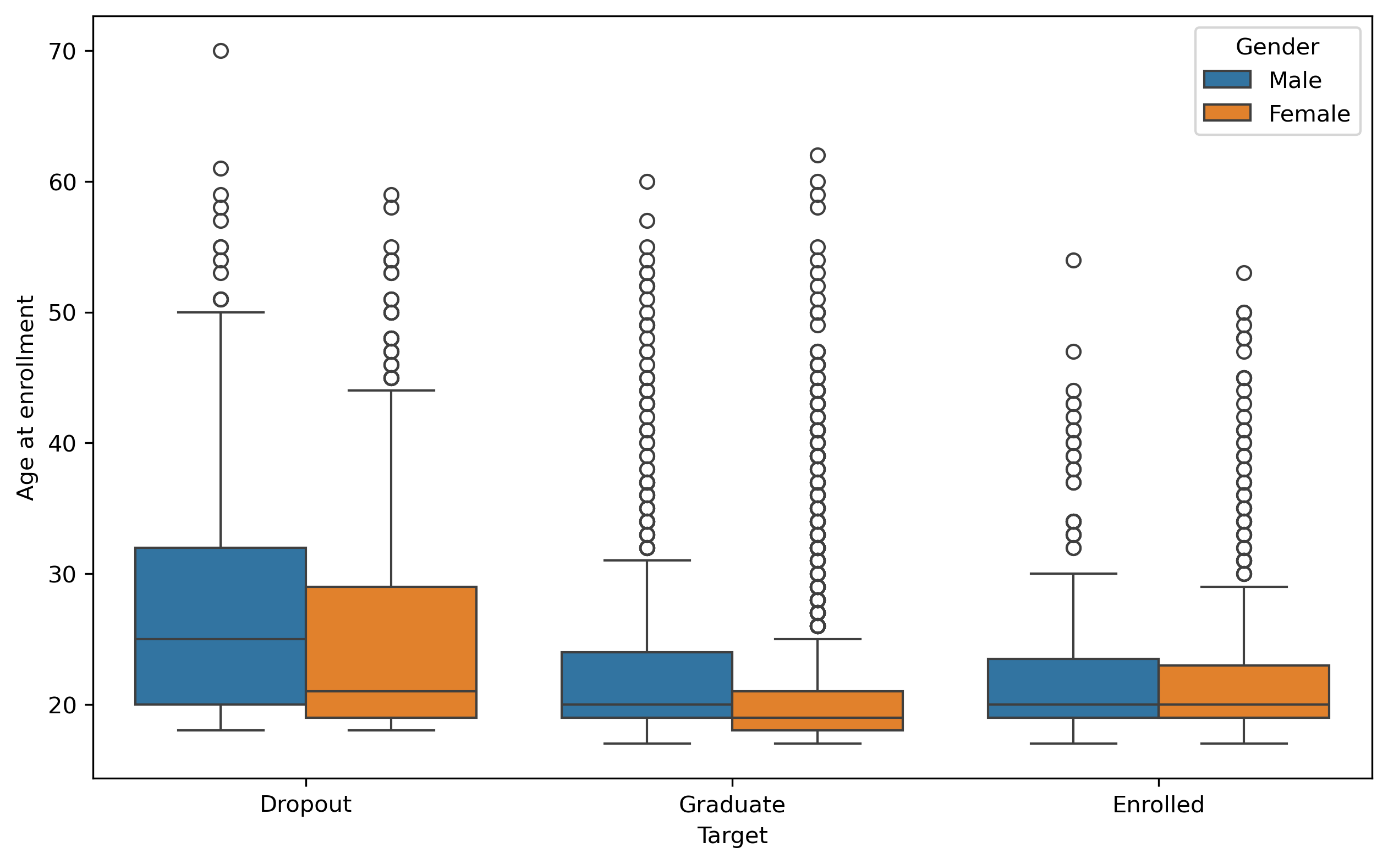
Observations

* Although older students tend to enroll in more classes and complete more evaluations, this does not translate into better outcomes, as both their average grades and number of approved units tend to decrease with age.
* The increasing number of evaluations taken by older students may indicate greater effort or course load, but the decreasing success rate suggests they may struggle more with academic performance, possibly due to external responsibilities or time away from formal education.



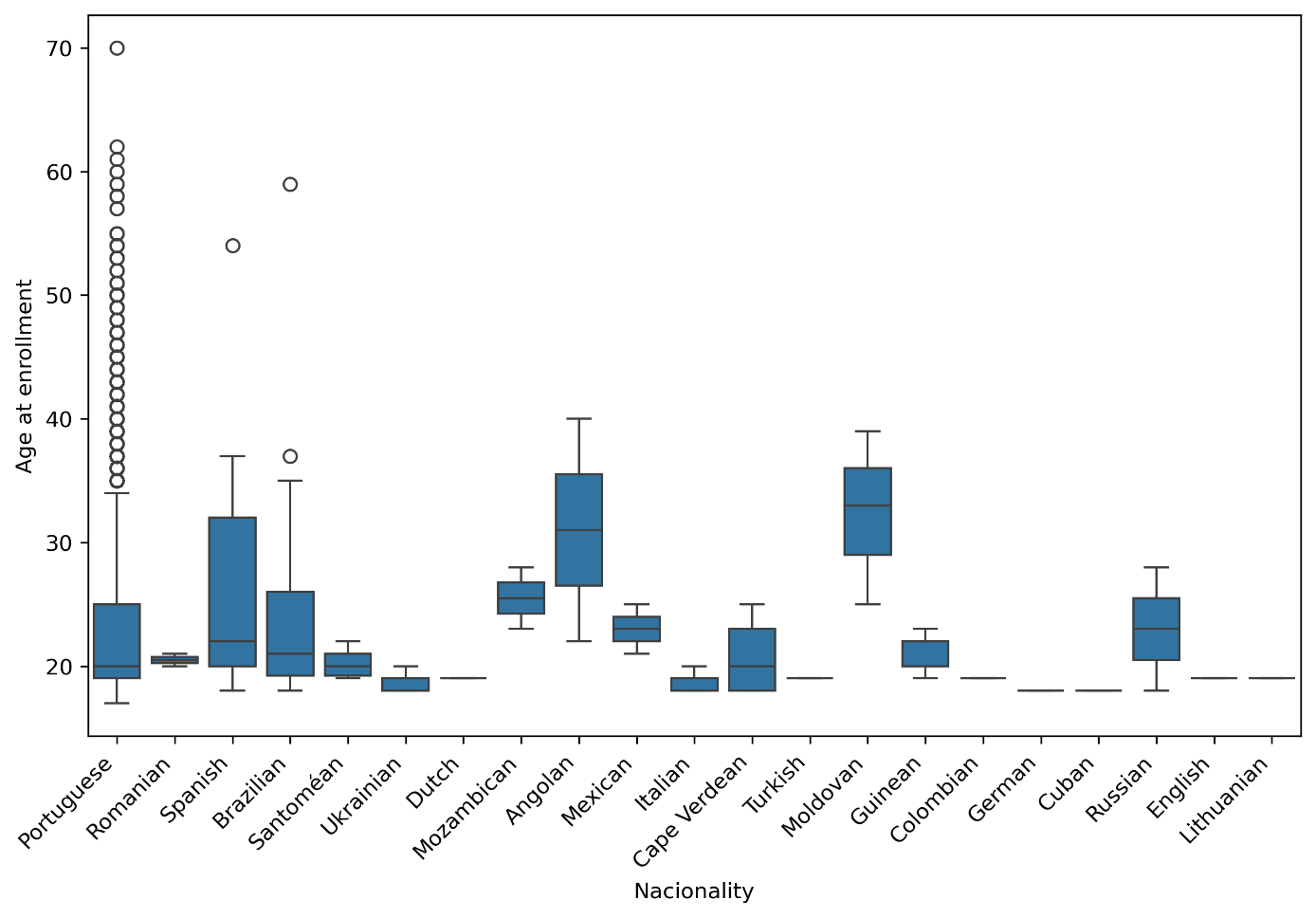
Observations

* Students who drop out tend to have a wider age distribution, including a significant number of older students, compared to those who graduate or remain enrolled.
* The graduate and enrolled groups show a much narrower age distribution, centered around 18–22 years, suggesting that younger students are more likely to succeed or persist in their studies.
* The increased presence of older students among dropouts supports the earlier findings that age may negatively affect academic outcomes.



Observations

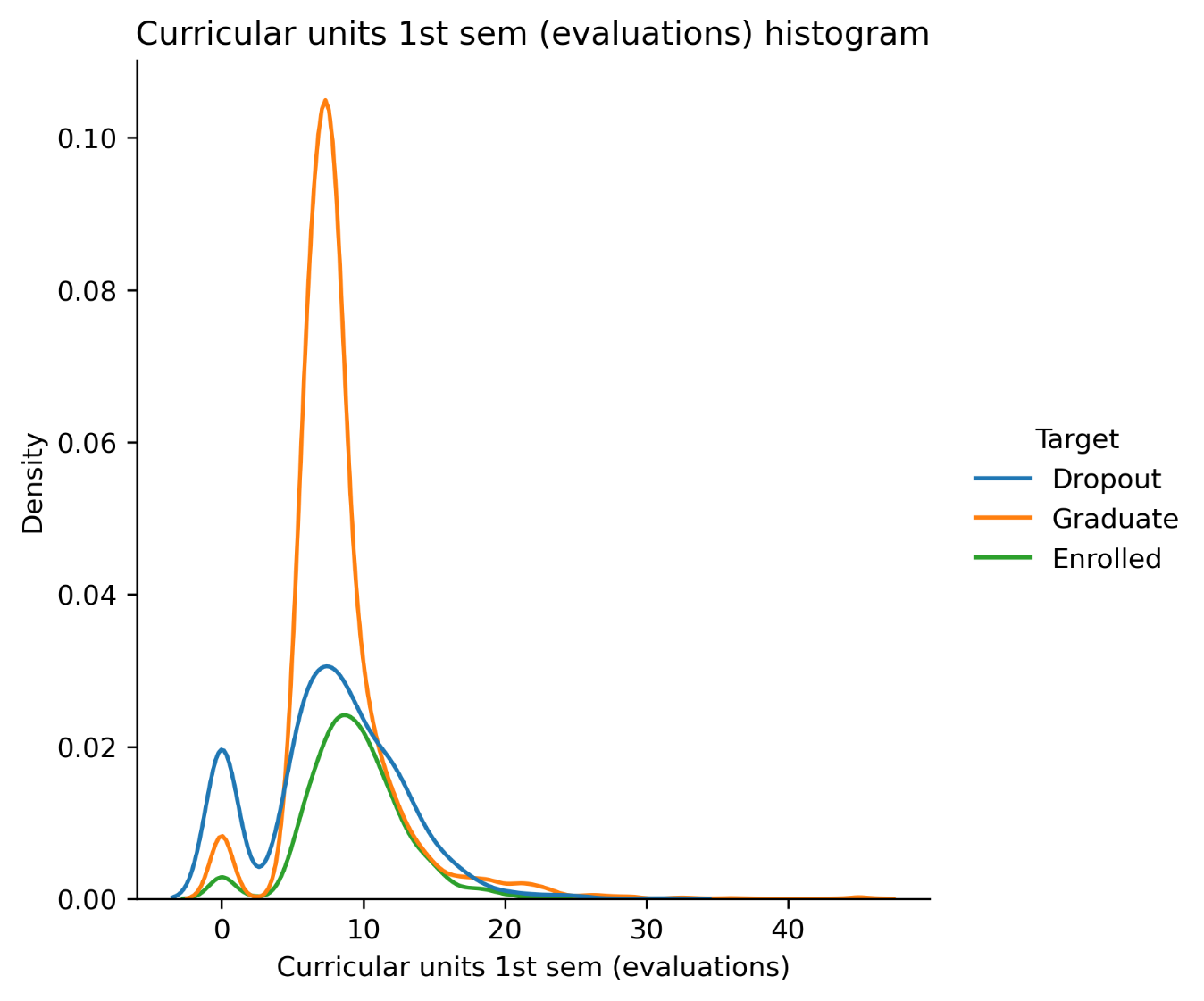
* Among students who drop out, males are on average older than females, and also show a wider spread of ages, including more older outliers.
* In the graduate and enrolled groups, both genders tend to enroll at a younger and more similar age, with distributions mostly centered around the early 20s.
* The age gap between genders is most noticeable among dropouts, suggesting that older male students may be at higher risk of not completing their studies.

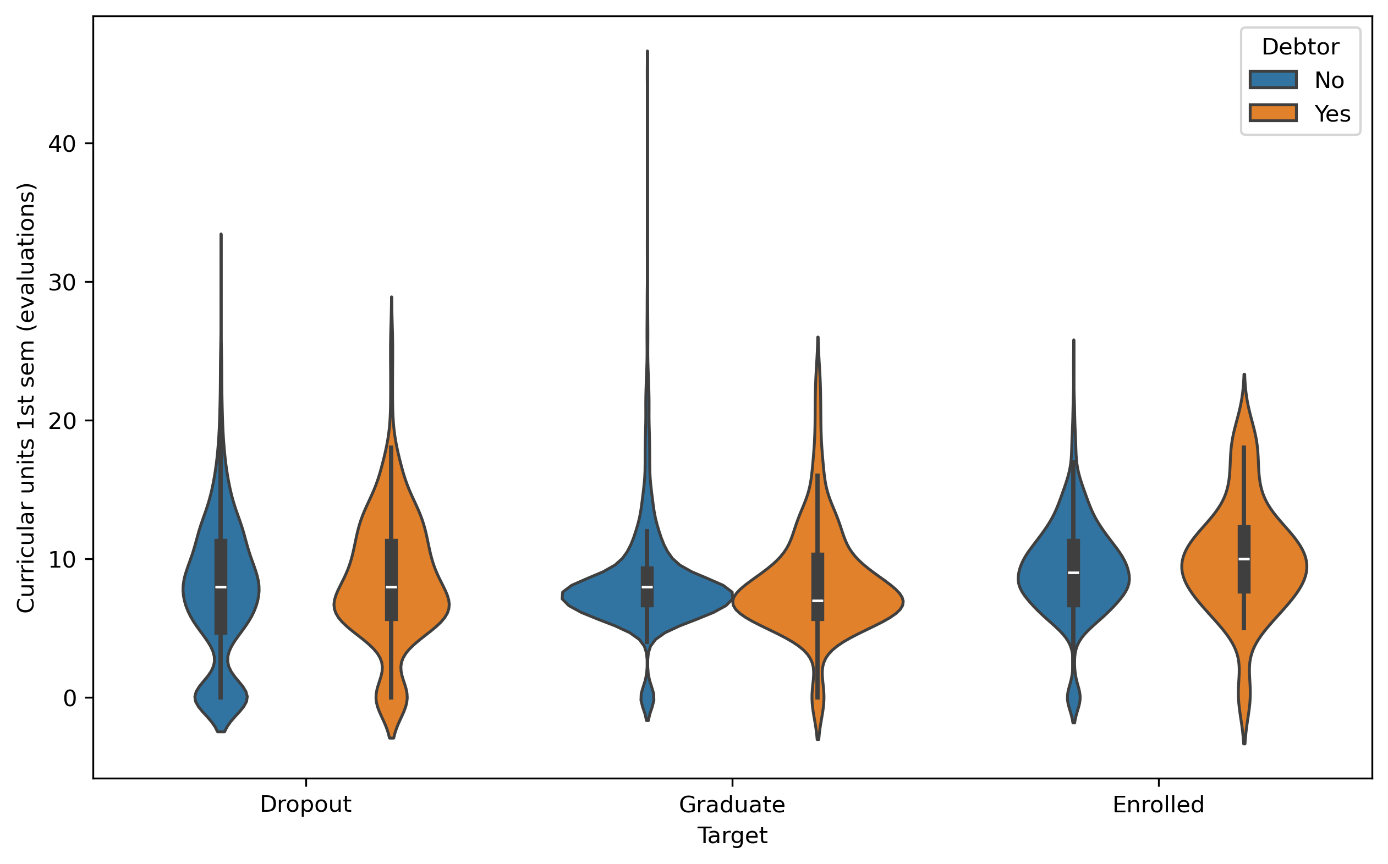


Observations

* Portuguese students show the widest range in age at enrollment, including the largest number of middle-aged and elderly students.
* Students from Brazil, Spain, Angola, and Moldova tend to enroll at older ages, with many having median enrollment ages in the late 20s or early 30s.
* In contrast, nationalities such as Romanian, Santomean, Ukrainian, and Cape Verdean display more compact age distributions, typically centered around early adulthood (18–22 years).
* These patterns suggest potential cultural, economic, or immigration-related factors influencing the timing of higher education enrollment across different nationalities.

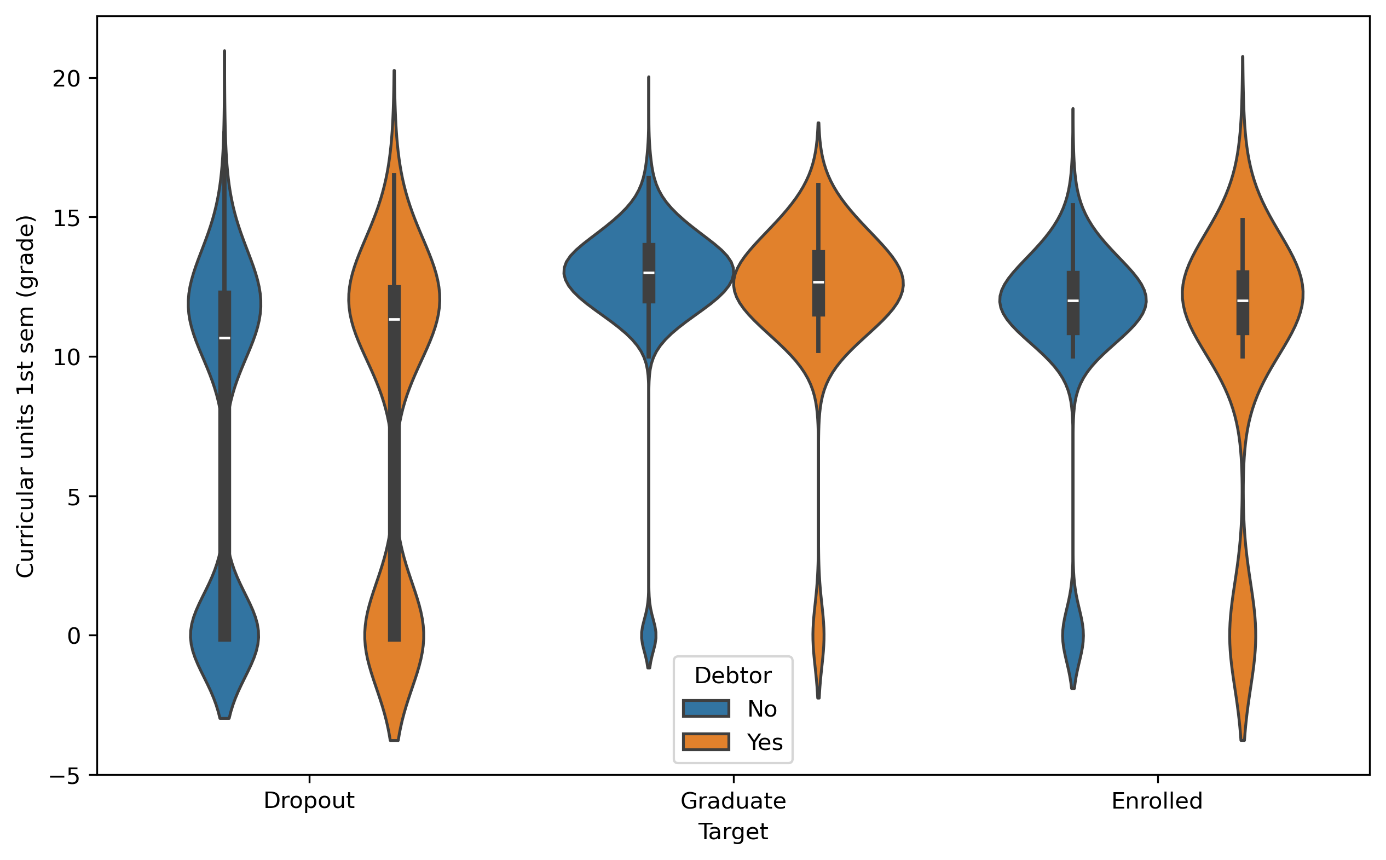
**1st semester**

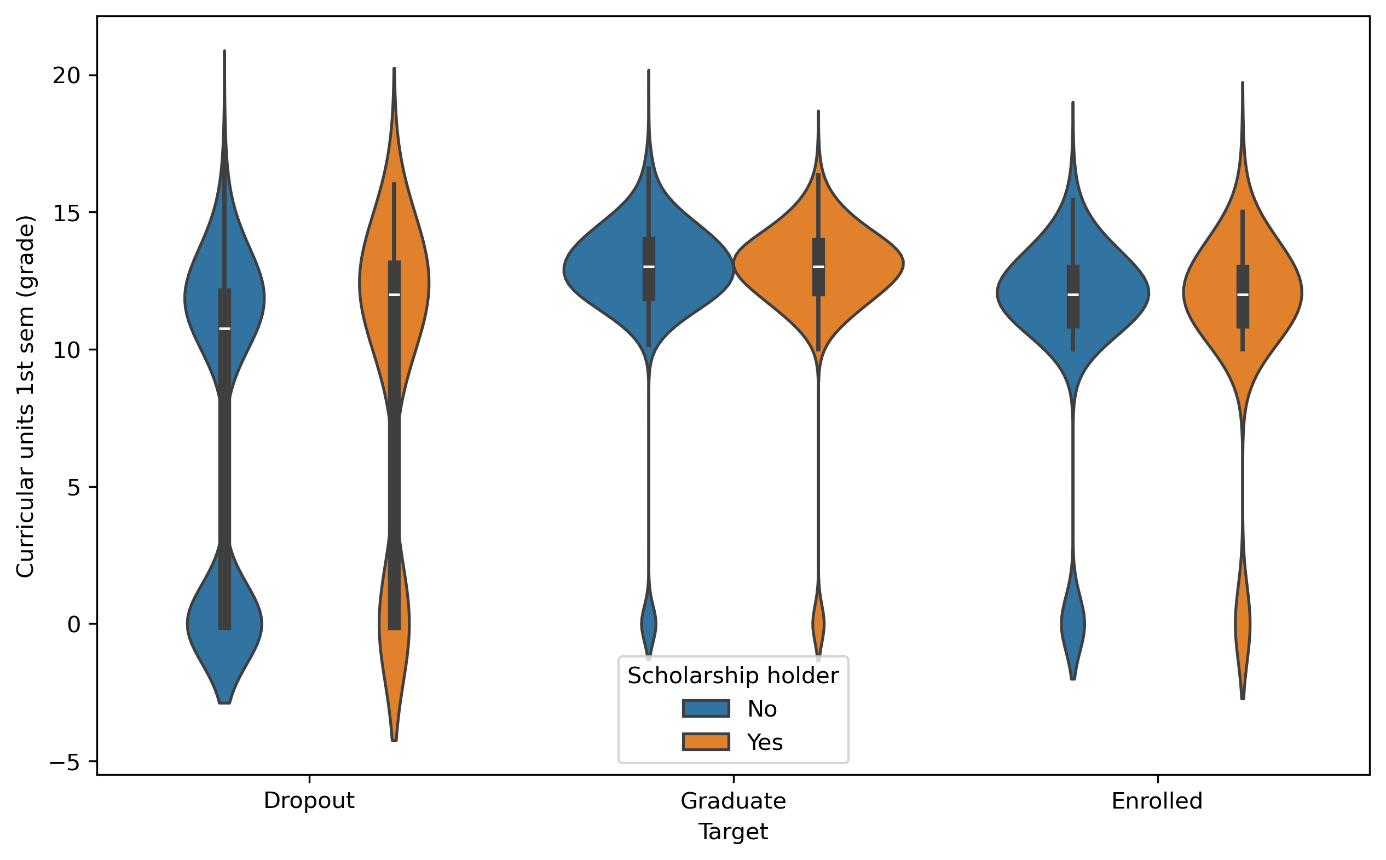
****

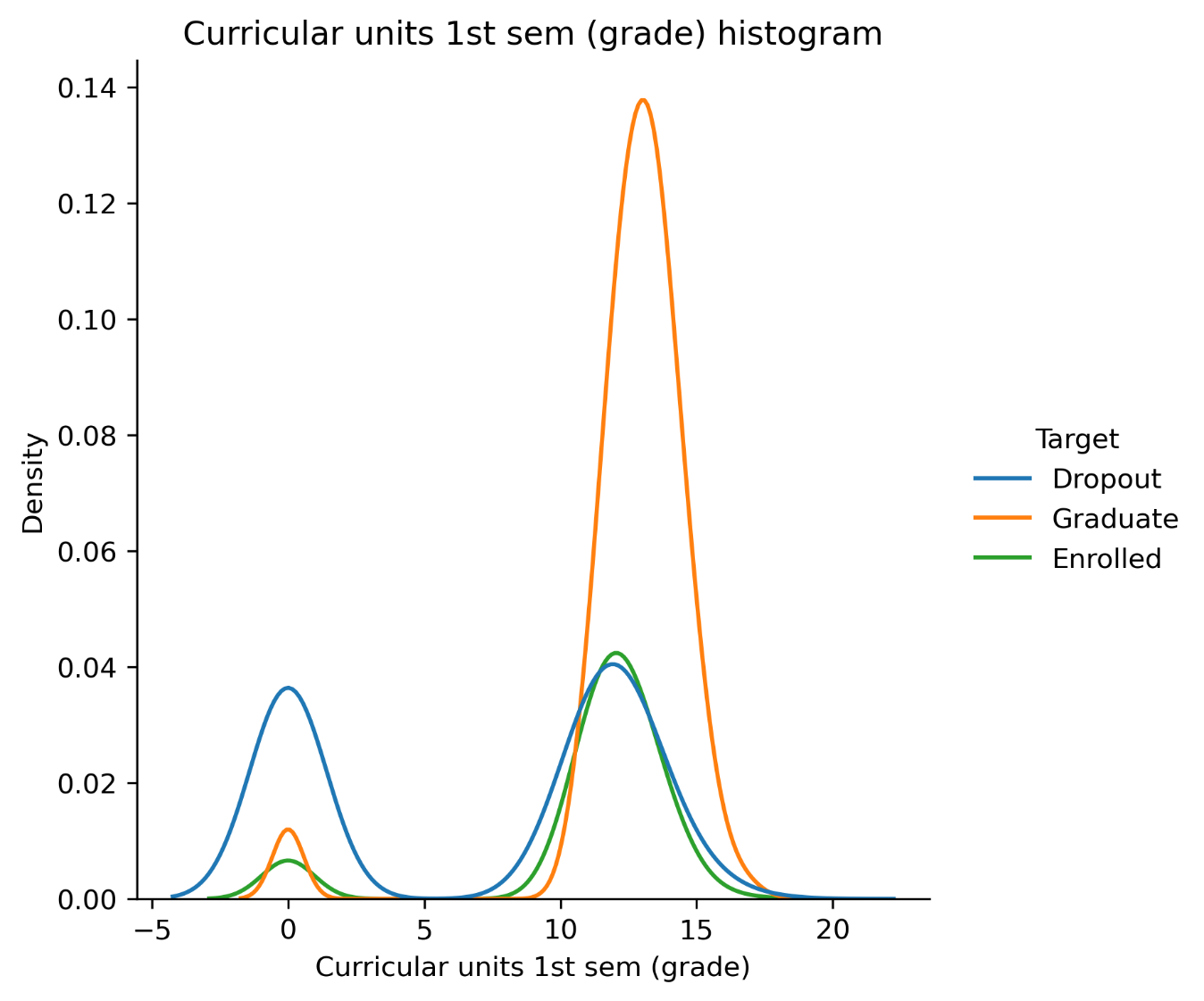
****

Observations

* Dropout students often have zero or very low numbers of evaluations during the first semester, indicating disengagement early in the academic process.
* The histogram shows a sharp peak around 7–9 evaluations for graduate students, suggesting they typically complete most of the expected assessments.
* Debtor students (those with financial or academic liabilities) tend to attempt more evaluations on average compared to non-debtors, possibly reflecting pressure to prove academic progress or fulfill specific requirements.

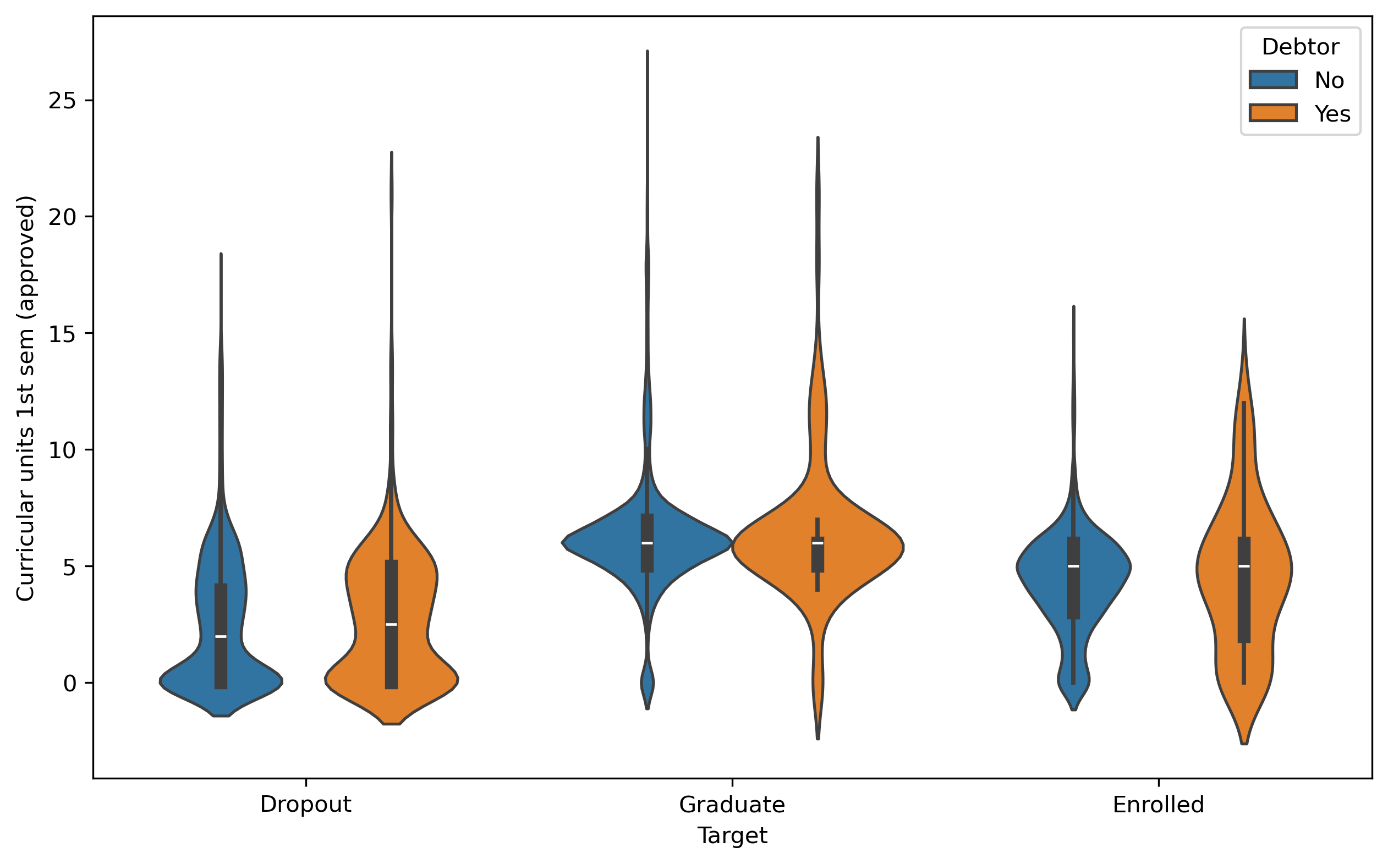


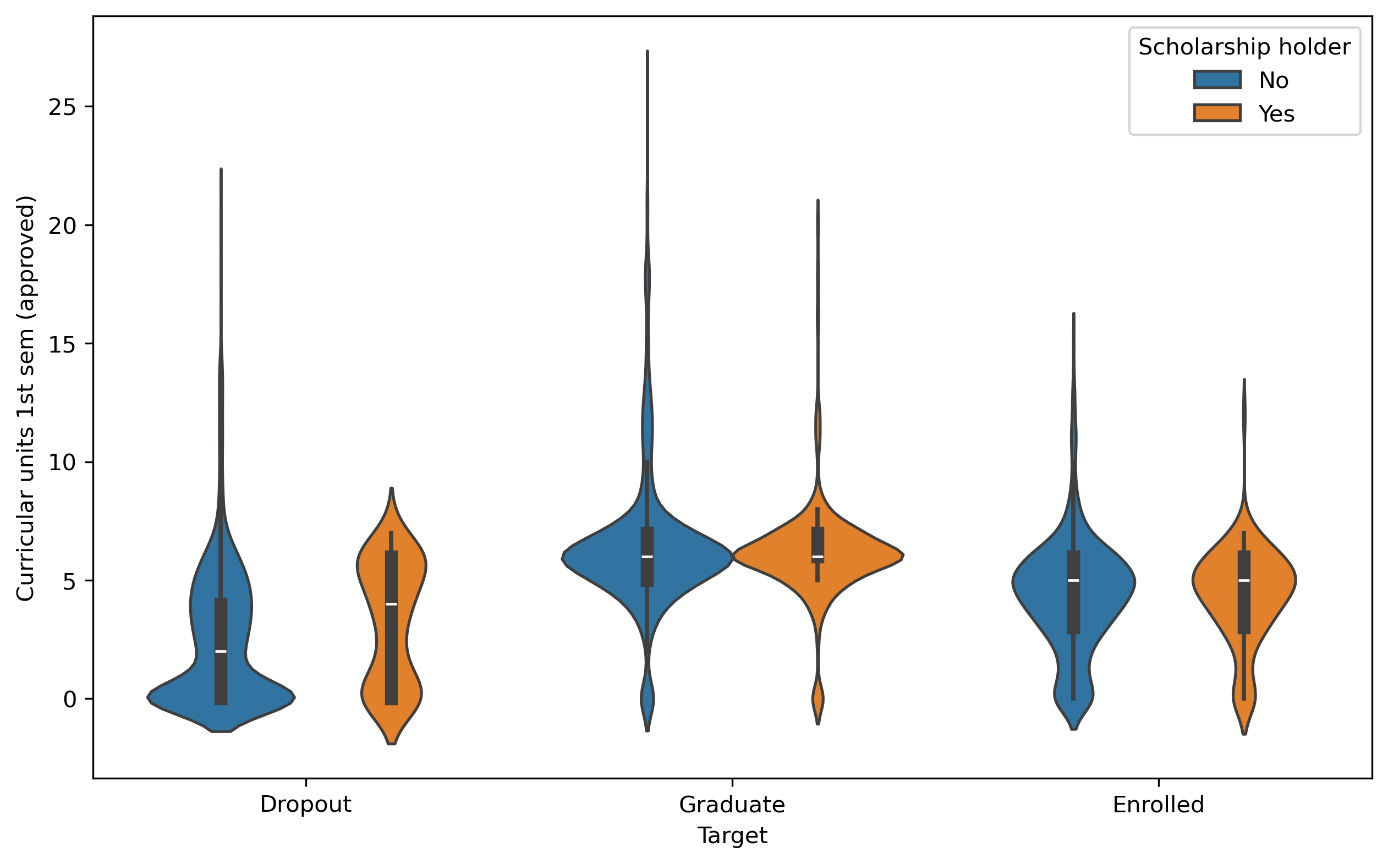


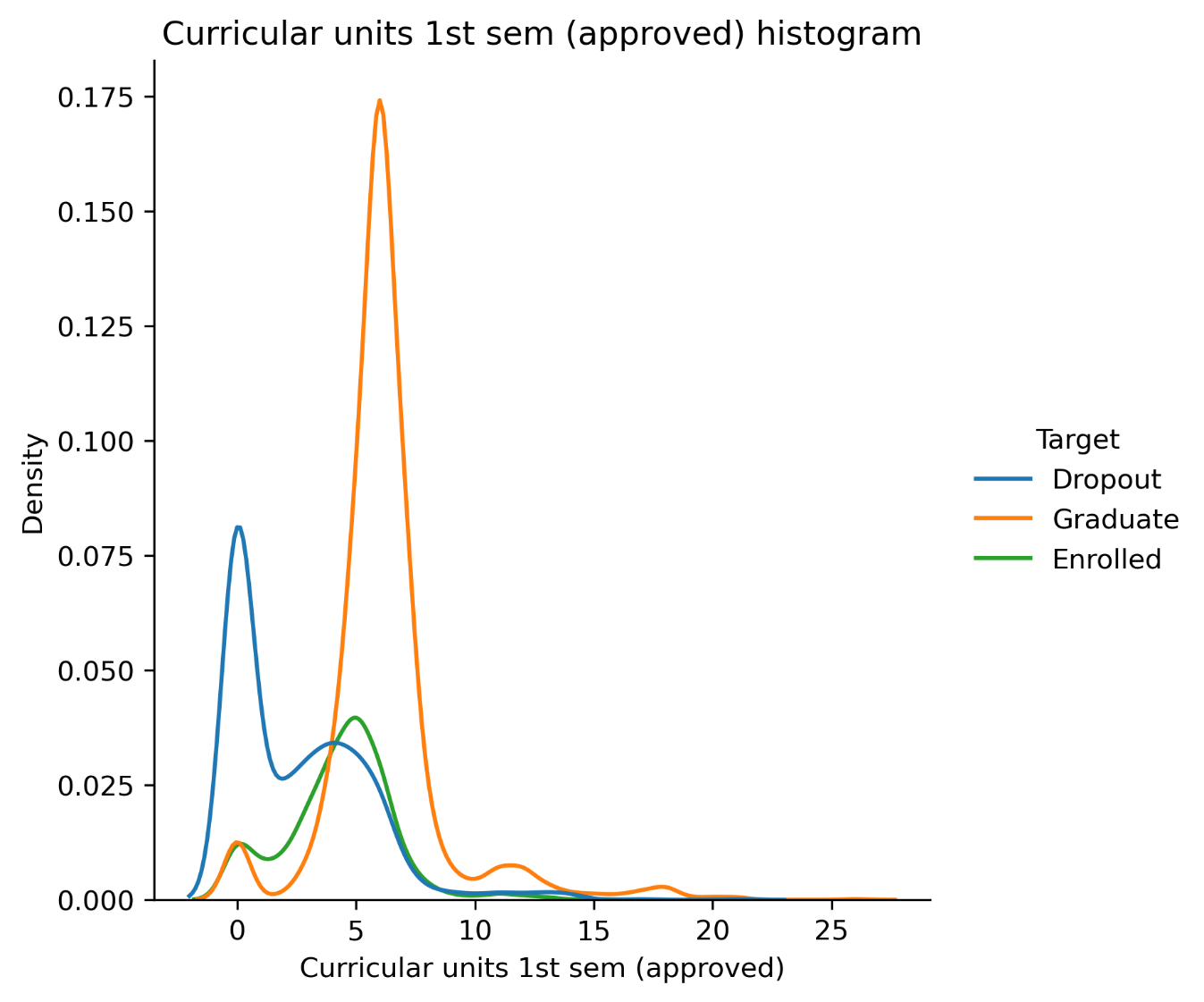


Observations

* A significant number of dropout students receive extremely low grades in their first semester, often close to zero, as clearly seen in the histogram.
* In contrast, graduate students tend to have high and consistent grades, typically centered around 13–14, with much lower variance.
* Debtor students and scholarship holders show more consistent and higher grades than their counterparts, especially among graduates and enrolled students. This may indicate that having financial or academic responsibility increases motivation or accountability.
* Among dropouts, students who were neither debtors nor scholarship holders tend to perform the worst — suggesting that external motivation may play a critical role in early academic success.
* Grade distributions for enrolled students are generally between those of dropouts and graduates, showing moderate performance and variation.



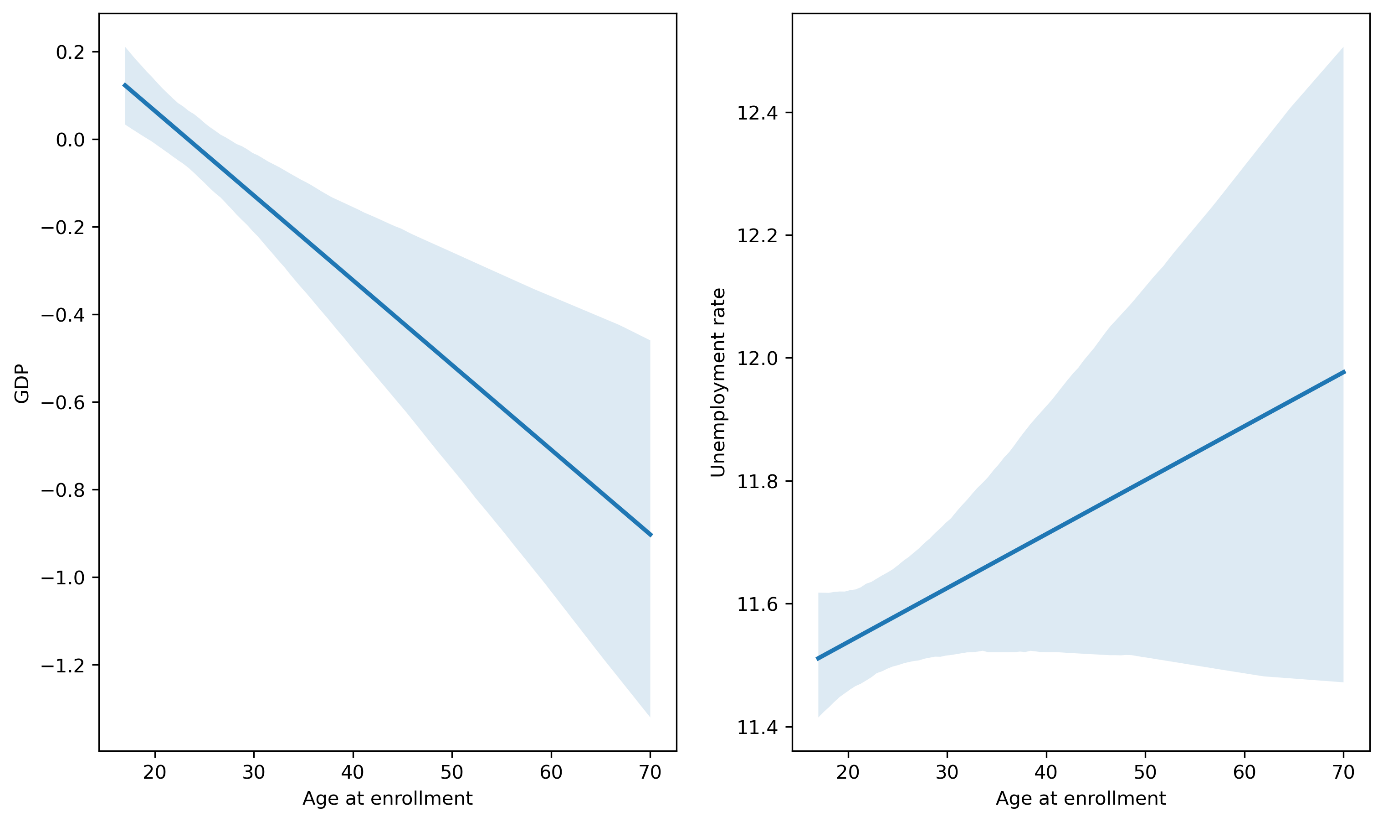




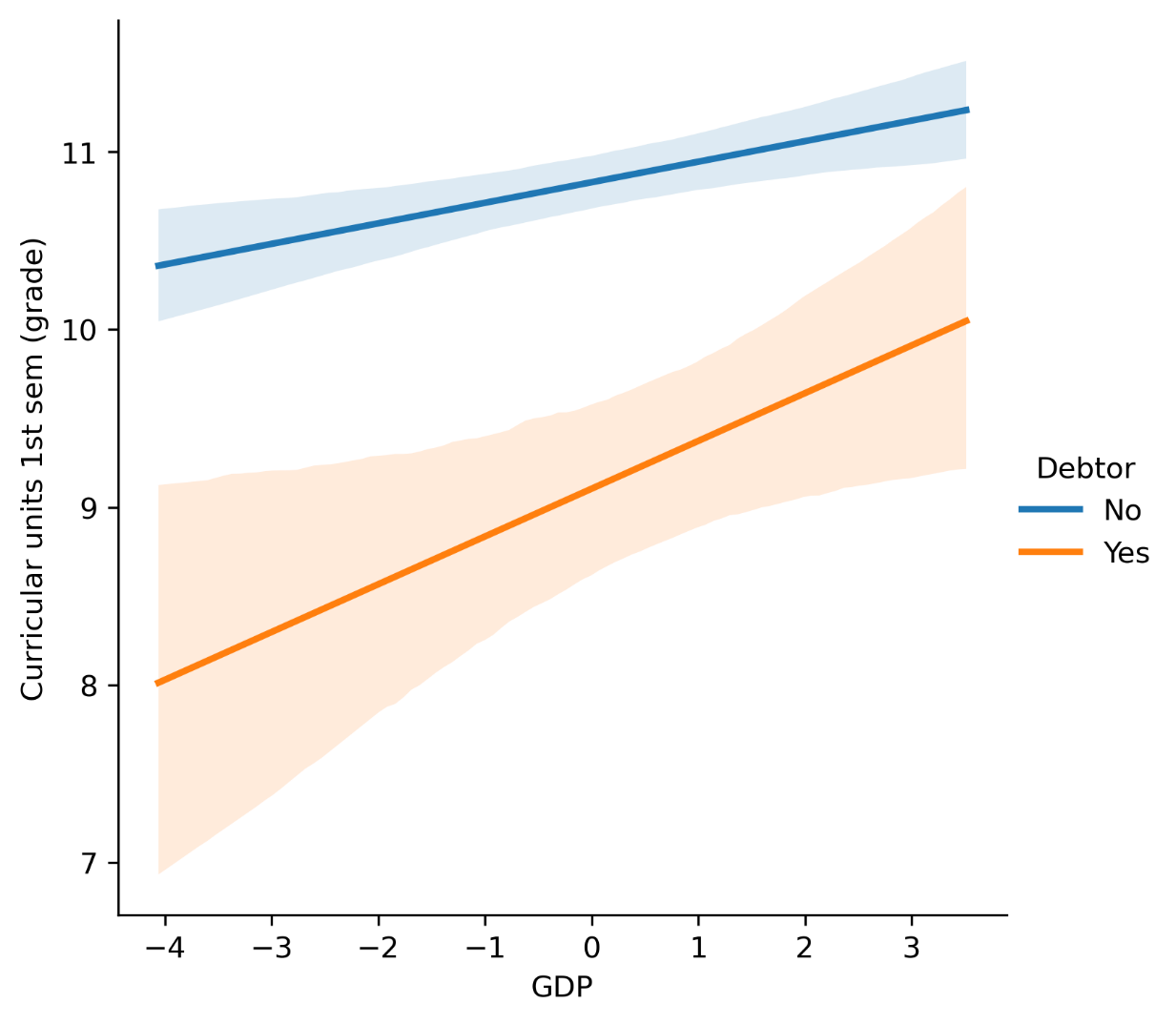
Observations

* Dropout students typically pass very few units in the first semester—most often between 0 and 2—highlighting a critical early academic struggle.
* The histogram confirms that graduates consistently pass the most units, typically around 6 or more, while enrolled students fall in between dropouts and graduates.
* Among dropouts, scholarship holders tend to perform slightly better than those without scholarships, indicating that financial support may buffer against failure.
* Debt status appears to have minimal impact on approval rates during the 1st semester, as distributions between debtor and non-debtor students are similar across all outcomes.

**Economic**

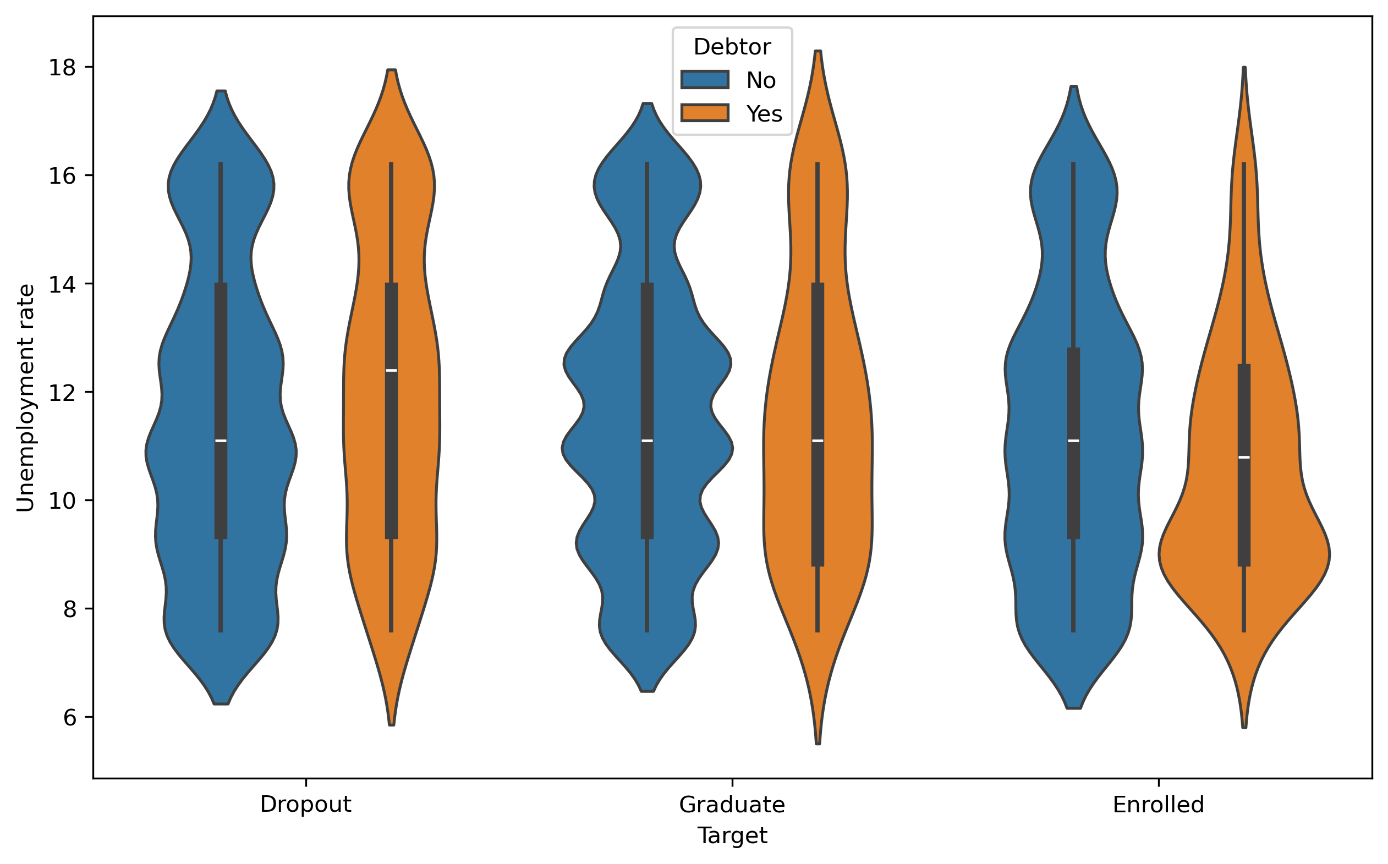
****

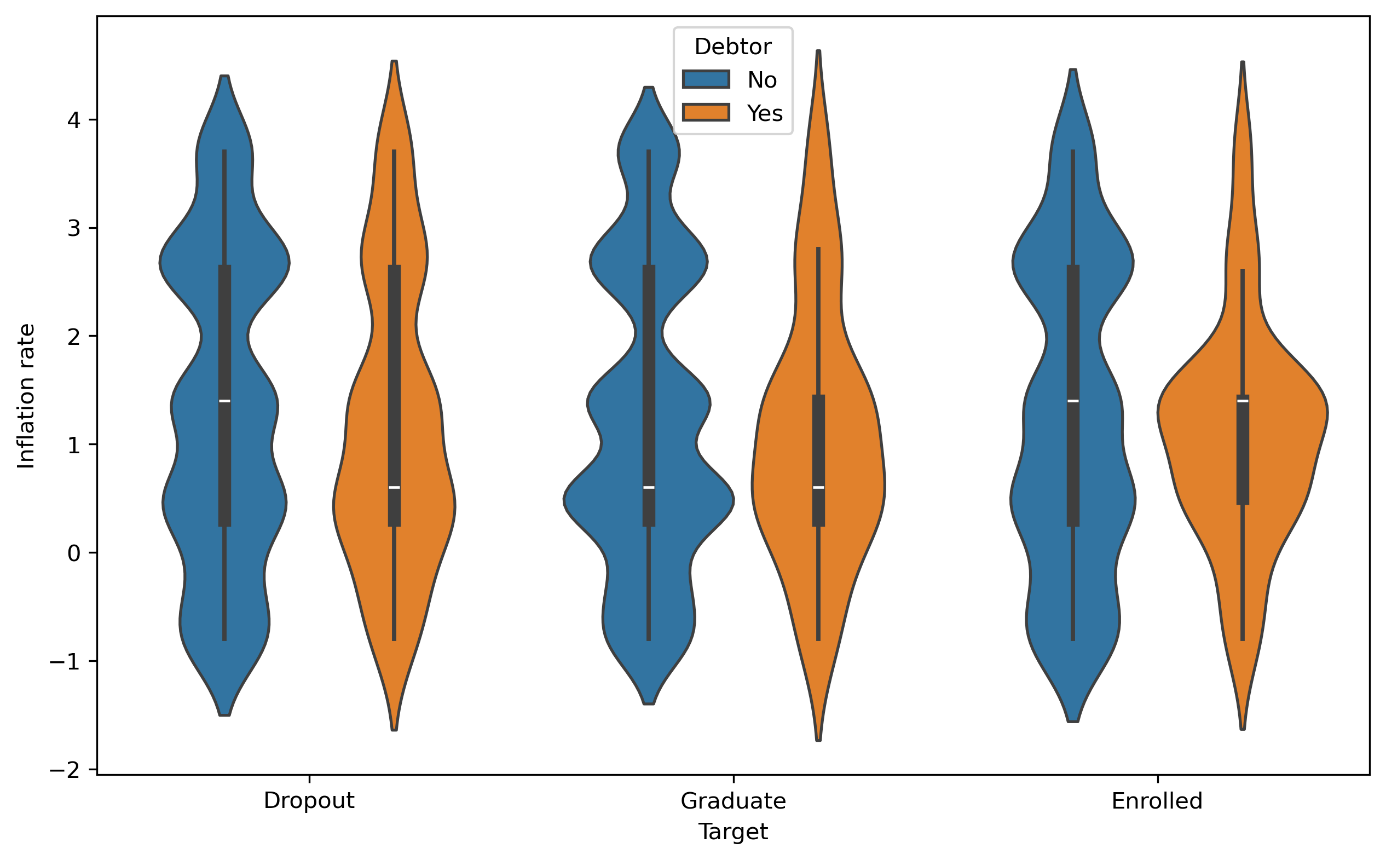
* The older the student, the more likely they enrolled during periods of lower GDP, suggesting that economic downturns may motivate adult learners to pursue education or reskilling.
* Similarly, there is a positive correlation between age at enrollment and unemployment rate, indicating that middle-aged and elderly students are more likely to enroll during times of higher unemployment.
* These patterns imply that non-traditional students might seek education as a response to unfavorable job market conditions, using academic programs as a way to reorient their careers or wait out economic hardship.

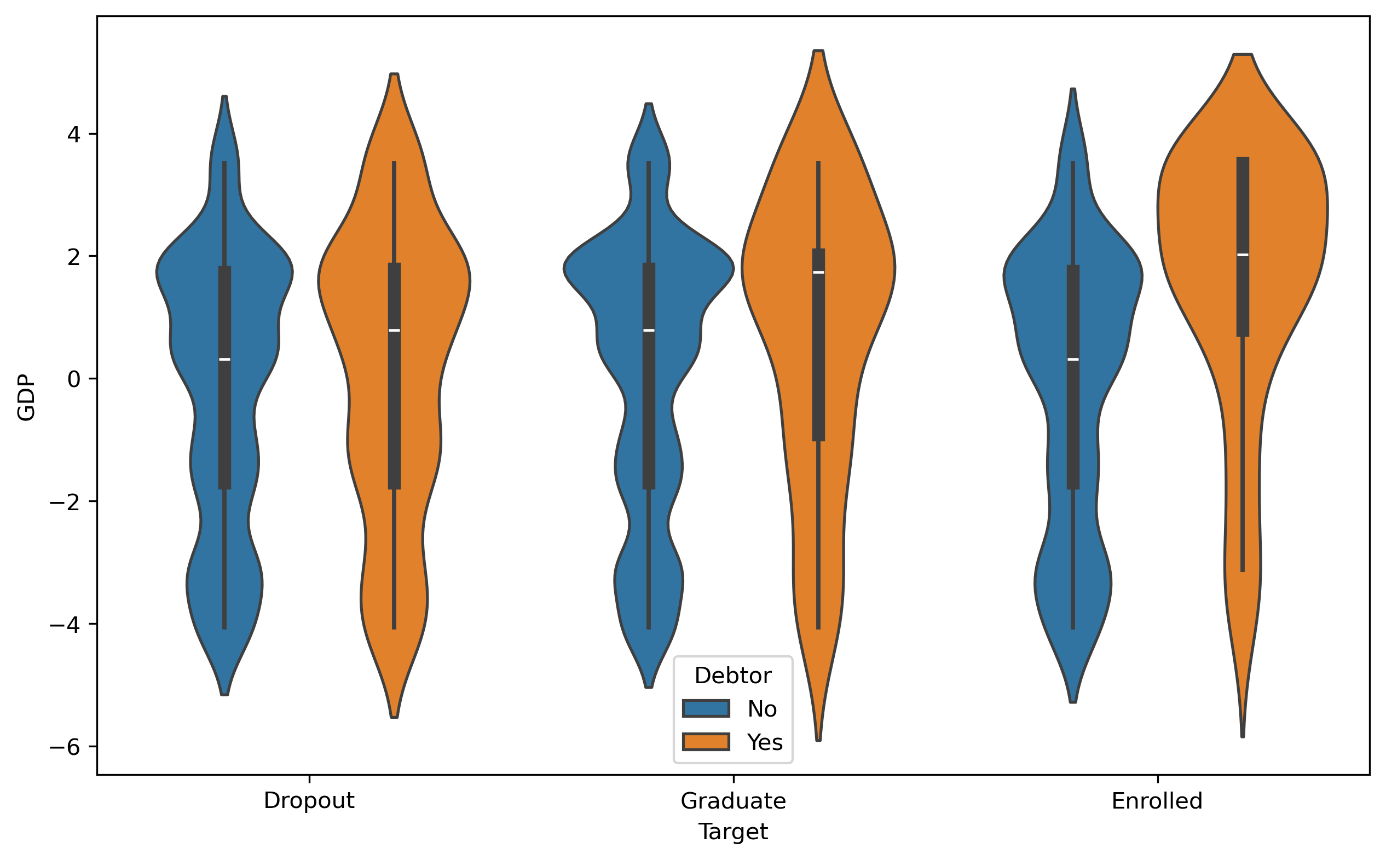


Observations

* Students tend to achieve better grades during periods of economic growth, as indicated by the positive correlation between GDP and 1st semester grades.
* Non-debtor students consistently perform better than debtor students across all economic conditions, though both groups benefit from a rising GDP.
* The performance gap between debtor and non-debtor students narrows as GDP increases, suggesting that economic stability may help disadvantaged students catch up.
* In times of economic decline (negative GDP), debtor students are especially vulnerable, with their grades dropping significantly more than those of their non-debtor peers.





****

Observations

* Students who are debtors are more likely to have enrolled during times of economic growth — indicated by the higher GDP values among debtors across all groups (dropout, graduate, enrolled).
* The unemployment rate is lower for debtor students, especially in the enrolled group, suggesting that students may accumulate debt during stable job market periods, potentially feeling safer to take financial risks.
* A similar trend is seen with inflation: debtor students typically enrolled during low-inflation periods, which could signal greater financial confidence or loan availability.
* These patterns suggest that favorable macroeconomic conditions (higher GDP, lower unemployment and inflation) are associated with increased student indebtedness, likely due to greater accessibility to credit and perceived economic safety.

**Part 2**

**Data portioning**

In each notebook the dataset is divided into the following sets

* training set – 64%
* validation set - 16%
* test set - 20%

**Model training and prediction using a pipeline**

Following column transformer was used:

* for numerical features:
  + replacement of missing values with the median
  + standardization
* for categorical features:
  + replacement of missing values with the most frequent category
  + one-hot encoding

The following models from the scikit-learn libarary were used:

1. Logistic Regression (LogisticRegression)
2. Decision Tree Classifier (DecisionTreeClassifier)
3. Support Vector Machine (SVC)

Each model was trained and evaluated five times for accuracy on the training, validation, and test datasets.

1. Logistic Regression (LogisticRegression)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 0.8100 | 0.8093 | 0.8100 | 0.8191 | 0.8093 |
| Validation | 0.7740 | 0.7797 | 0.7938 | 0.7429 | 0.7768 |
| Test | 0.7797 | 0.7864 | 0.7763 | 0.7853 | 0.7729 |

1. Decision Tree Classifier (DecisionTreeClassifier)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 0.8001 | 0.8117 | 0.8025 | 0.8011 | 0.8061 |
| Validation | 0.7373 | 0.7218 | 0.7641 | 0.7302 | 0.7246 |
| Test | 0.7458 | 0.7435 | 0.7627 | 0.7277 | 0.7401 |

1. Support Vector Machine (SVC)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 0.8365 | 0.8400 | 0.8410 | 0.8365 | 0.8301 |
| Validation | 0.7825 | 0.7458 | 0.7458 | 0.7571 | 0.7782 |
| Test | 0.7808 | 0.7842 | 0.7819 | 0.7785 | 0.7842 |

The Logistic Regression and SVC models achieve similar accuracies across training, validation, and test sets. On the other hand, the Decision Tree Classifier achieves worse results, which might be the matter of parametrization.

**Closed-form Linear Regression**

Linear regression has the following equation:

where

y – vector of observed values

X – matrix of observations (rows) and features (columns)

w – vector of weights

Then the normal equation to compute the optimal weights *w* directly:

**Linear Regression with Gradient Descent**

This implementation iteratively minimizes the MSE over the training data to optimize the weights.

The linear regression model assumes

The weights are iteratively updated using

Where the gradient for the const function is given by

**Linear Regression implementations**

Each linear regression implementation was trained and evaluated five times for mean squared error on the training, validation, and test datasets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 136.4140 | 135.5521 | 136.2144 | 137.0914 | 134.5934 |
| Validation | 128.7778 | 131.9047 | 130.4448 | 126.2236 | 135.9268 |
| Test | 131.2311 | 130.6790 | 130.6887 | 130.8080 | 130.3862 |

1. Closed-form Linear Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 136.5241 | 135.7230 | 136.3940 | 137.2926 | 134.7437 |
| Validation | 128.9442 | 131.9764 | 130.1461 | 126.5491 | 136.1407 |
| Test | 131.4243 | 130.6763 | 130.1332 | 130.7191 | 130.5336 |

1. Linear Regression with Gradient Descent

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 136.4140 | 135.5521 | 136.2144 | 137.0914 | 134.5934 |
| Validation | 128.7778 | 131.9047 | 130.4448 | 126.2236 | 135.9268 |
| Test | 131.2311 | 130.6190 | 130.6887 | 130.8080 | 130.3862 |

1. Sklearn Linear Regression

All three implementations produce comparable results. However, the gradient descent approach turns out to be slightly less effective, which is expected due to its iterative nature and reliance on stochasticity. On the other hand, both the closed-form solution and scikit-learn's implementation yield identical results.

**Logistic Regression with Gradient Descent**

This implementation iteratively minimizes the binary cross-entropy loss over the training data to optimize the weights.

The logistic regression model assumes

where is the sigmoid function, X is the input matrix, and w is the weight vector.

The weights are iteratively updated using

Where the gradient for the const function is given by

Each logistic regression implementation was trained and evaluated five times for accuracy on the training, validation, and test datasets

1. Custom Logistic Regression with gradient descent

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 0.8199 | 0.8124 | 0.8209 | 0.8244 | 0.8223 |
| Validation | 0.7910 | 0.8023 | 0.7881 | 0.7528 | 0.7726 |
| Test | 0.7571 | 0.7684 | 0.7661 | 0.7650 | 0.7684 |

1. Sklearn Logistic Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 1 | 2 | 3 | 4 | 5 |
| Train | 0.7838 | 0.7838 | 0.7831 | 0.7817 | 0.7821 |
| Validation | 0.7655 | 0.7868 | 0.7655 | 0.7797 | 0.7952 |
| Test | 0.7537 | 0.7559 | 0.7514 | 0.7548 | 0.7571 |

The custom logic regression implementation with gradient descent provides slightly worse evaluations compared to the scikit-learn’s implementation, which might be the matter of parametrization.

**Linear Regression with PyTorch**

The model is built as a simple neural network with:

* a single linear layer (only one layer of neurons that perform linear transformation   
  y = Wx+ b)
* Mean Squared Error (MSE) as the loss function
* an optimizer such as Stochastic Gradient Descent (SGD)

The training loop consists of:

* forward pass through the model (returns output / prediction with the current weights)
* computation of the loss using a cost function (MSE)
* backward pass (backpropagation, gradient computation to determine how each parameter / weight contributed to the output / prediction)
* weight update via the optimizer (adjusts weights to reduce the loss)

Linear regression was trained and evaluated five times for mean squared error on the training, validation, and test datasets for GPU and CPU

1. CPU

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| Training time | 3.98 | 3.21 | 3.90 | 3.89 | 3.88 |
| Train | 133.4656 | 136.5765 | 137.6624 | 138.1828 | 138.0577 |
| Validation | 151.2813 | 138.6728 | 135.2426 | 132.1579 | 132.6512 |
| Test | 121.7871 | 123.2026 | 124.0619 | 122.4109 | 122.2585 |

1. GPU

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| Training time | 6.13 | 6.03 | 5.93 | 5.94 | 5.97 |
| Train | 133.5033 | 136.5796 | 137.6080 | 138.1756 | 138.0409 |
| Validation | 151.162277 | 138.6446 | 134.9460 | 132.2841 | 132.5421 |
| Test | 121.6101 | 123.3583 | 123.9044 | 122.5772 | 122.2799 |

The comparison between CPU and GPU execution reveals that both configurations achieve similar results in terms of MSE. The difference is negligible, confirming that the model behaves consistently regardless of the computational backend.

Interestingly, the GPU runtime is higher than the CPU runtime, which may seem counterintuitive at first. However, this can be explained by the overhead associated with data transfer between the CPU and GPU, which includes CUDA library setup or VRAM initialization. For small datasets and not complicated models like in this case, the GPU advantage is not fully utilized. Nevertheless, more complex computing problems are where GPU truly outperforms CPU.

**Part 3**

**Introduction**

The third and final part of the project focuses on optimizing previously developed machine learning models to enhance their generalization and performance.

1. **Cross validation**

In this notebook, K-Fold cross validation was applied with k = 3, which means the dataset was split into three equally sized folds. The model was trained on two folds and validated on the remaining one, repeating this process three times so that each fold served as the validation set once. This approach helps to reduce the bias associated with random train/test split and provides a more robust estimate of model performance.

**Sklearn Logistic Regression**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **1** | **2** | **3** |
| **Accuracy** | 0.7654 | 0.7749 | 0.7782 |
| **Precision** | 0.7539 | 0.7617 | 0.7609 |
| **Recall** | 0.7654 | 0.7749 | 0.7782 |

**Custom Implementation of Logistic Regression**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **1** | **2** | **3** |
| **Accuracy** | 0.7661 | 0.7539 | 0.7741 |
| **Precision** | 0.7461 | 0.7317 | 0.7552 |
| **Recall** | 0.7661 | 0.7539 | 0.7741 |

**Interpretation**

The results from the 3-fold cross-validation indicate consistent model performance across all folds for both implementations.

* For Sklearn Logistic Regression the accuracy varies slightly, which suggests stable behavior and good generalization of the model. Fold 3 yields the best performance across all metrics (accuracy, precision, and recall), suggesting that the data split in this fold aligns best with the model’s learning dynamics.
* The custom implementation of logistic regression also shows comparable stability.

The small variance in metrics across folds in both cases indicates that neither implementation suffers from high variance or overfitting. The custom model performs closely to the sklearn baseline, confirming its correctness and reliability.

1. **Fd**
2. **3dsfs**
3. **Fdfdsf**
4. **Fsdfa**
5. **Fsdafa**

**Summary**