DSC-530 TERM FINAL

Bresa McClatchy

March 2, 2025

STATISTICAL QUESTION

What factors are most associated with student dropout rates and academic success?

THE DATASET: PREDICT STUDENTS DROPOUT, ACADEMIC SUCCESS

Pulled from Kaggle, this dataset was created from a higher education institution related to students enrolled in different undergraduate degrees.

VARIABLES USED FOR ANALYSIS

Target: The classification used for whether a students has dropped out, enrolled, or graduated.

Age at Enrollment: The age of the student at the time of enrollment.

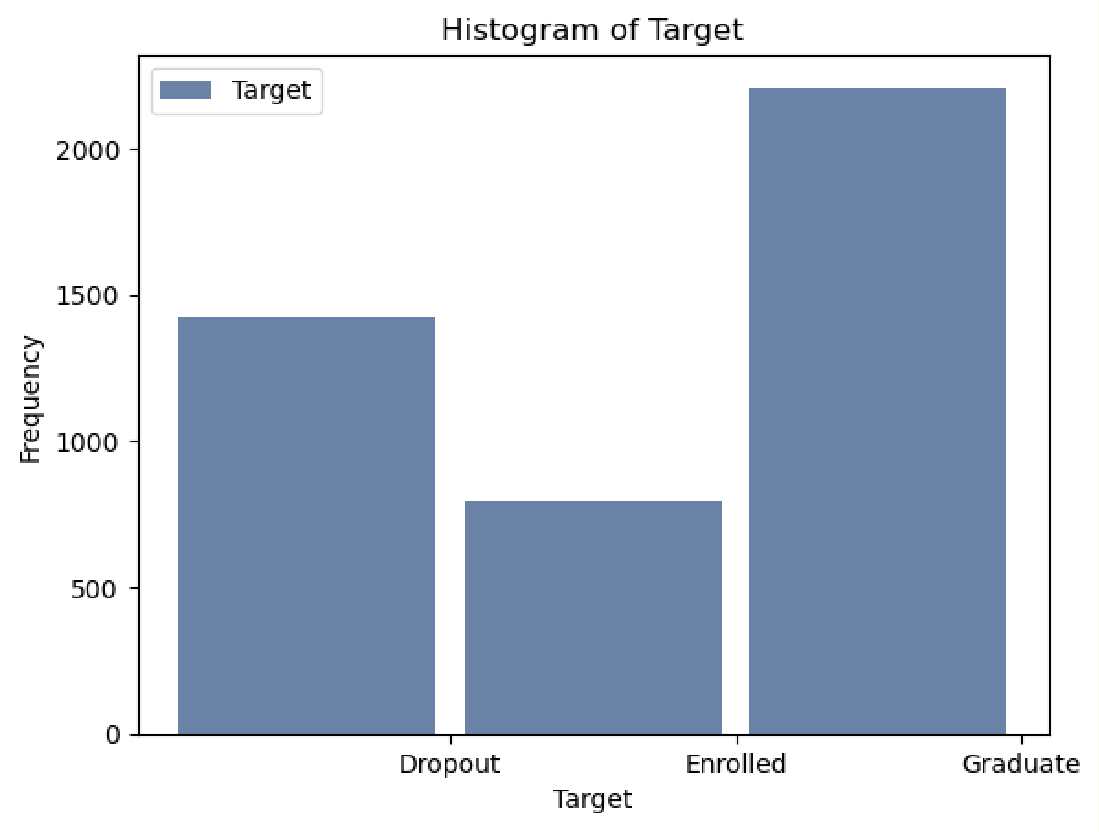
Scholarship Holder: Whether the student holds a scholarship.

Curricular Units 1st Semester: Grade average for 1st semester.

Curricular Units 2nd Semester: Grade average for 2nd semester.

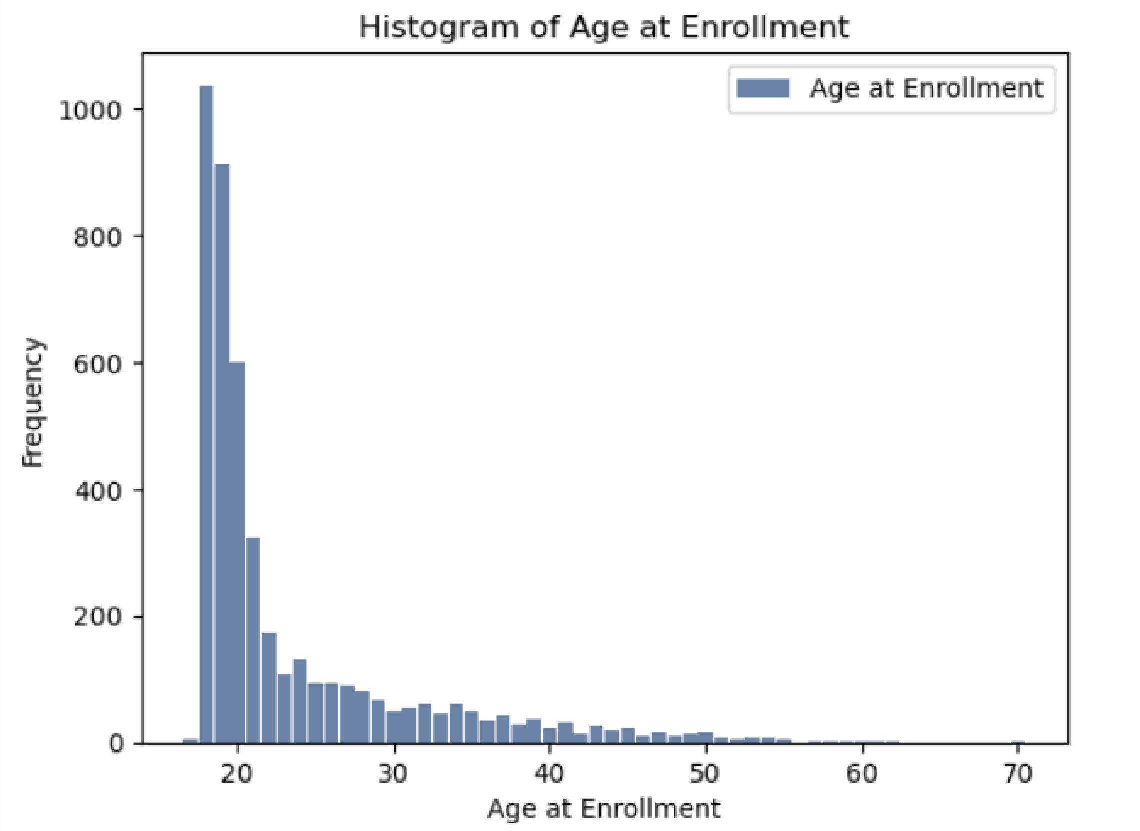
ANALYSIS OF DATA

TARGET (DROPOUT/ENROLLED/GRADUATE)



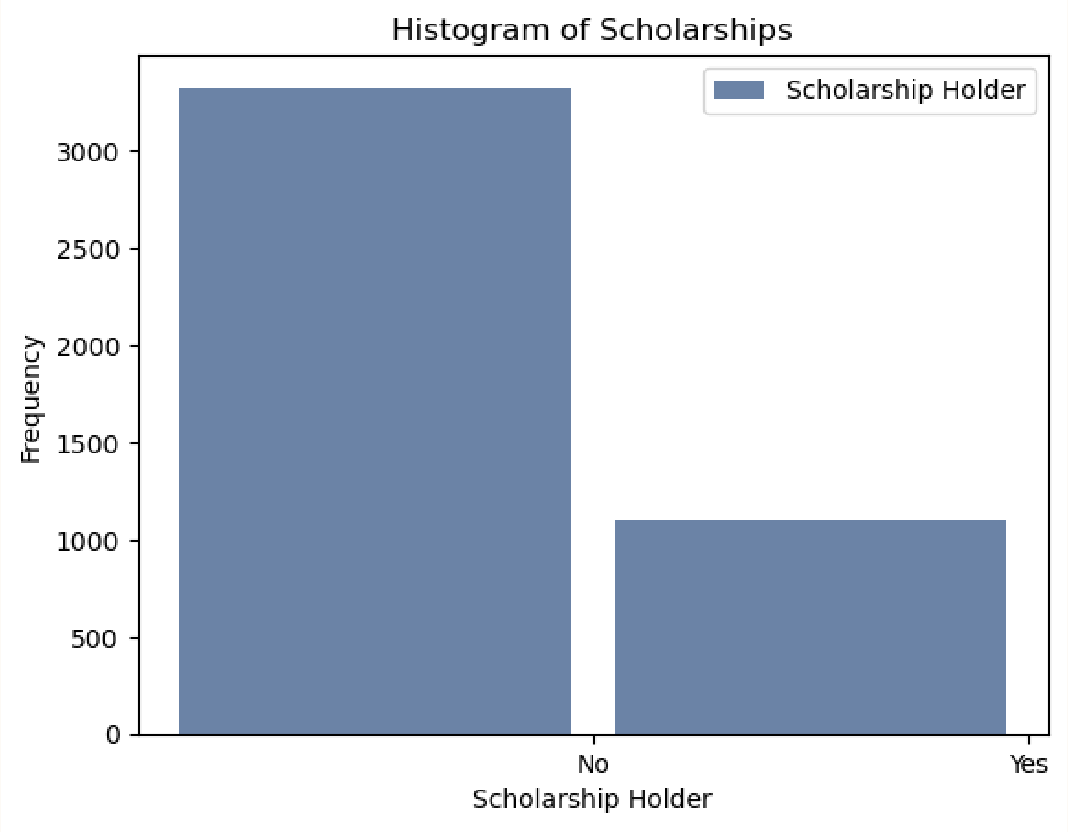
- Target is a categorical variable, therefore no outliers.
- Mean: 1.18
- Mode: 2
- Spread:
  - Standard Deviation: 0.89
  - Range: 2
- Tails:
  - Skewness: -0.36
  - Kurtosis: -1.64

AGE AT ENROLLMENT



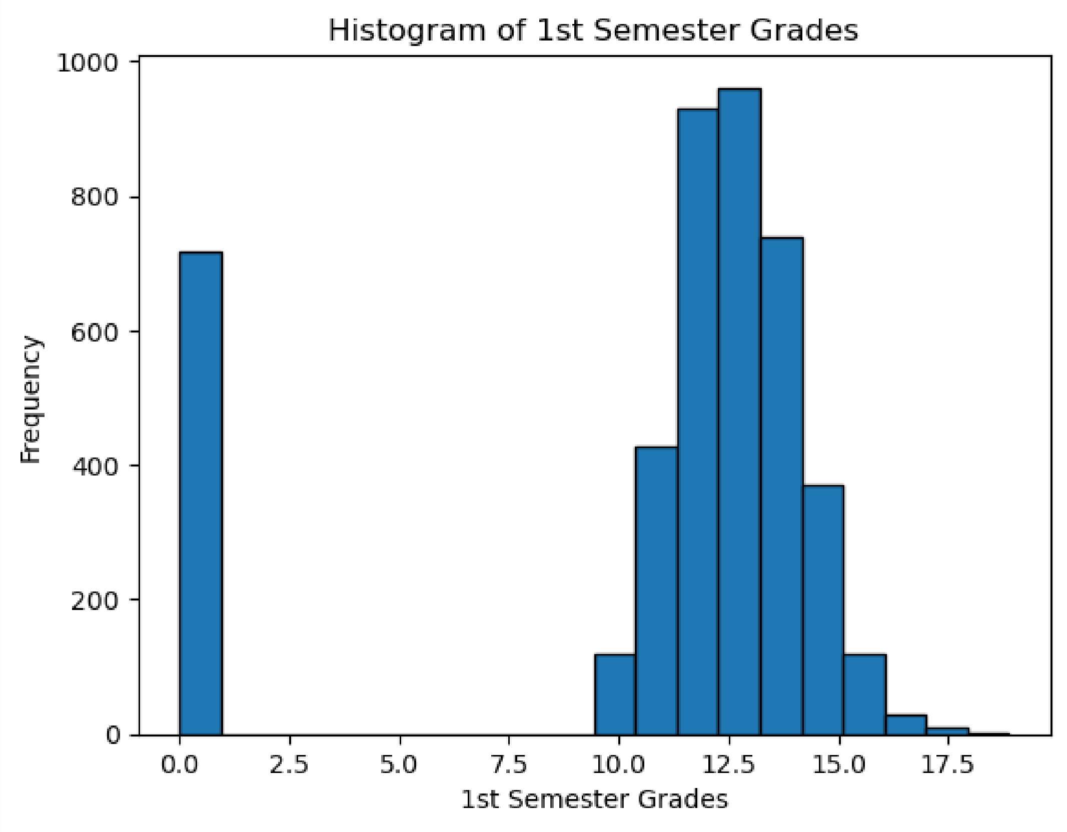
- Outliers in this instance are accurate reports of rare events, such as those who are enrolled over the age of 40. Normally, investigate, but likely correct, which might skew the data.
- Mean: 23.27
- Mode: 18
- Spread:
  - Standard Deviation: 7.59
  - Range: 53
- Skew: 2.05

SCHOLARSHIP HOLDER



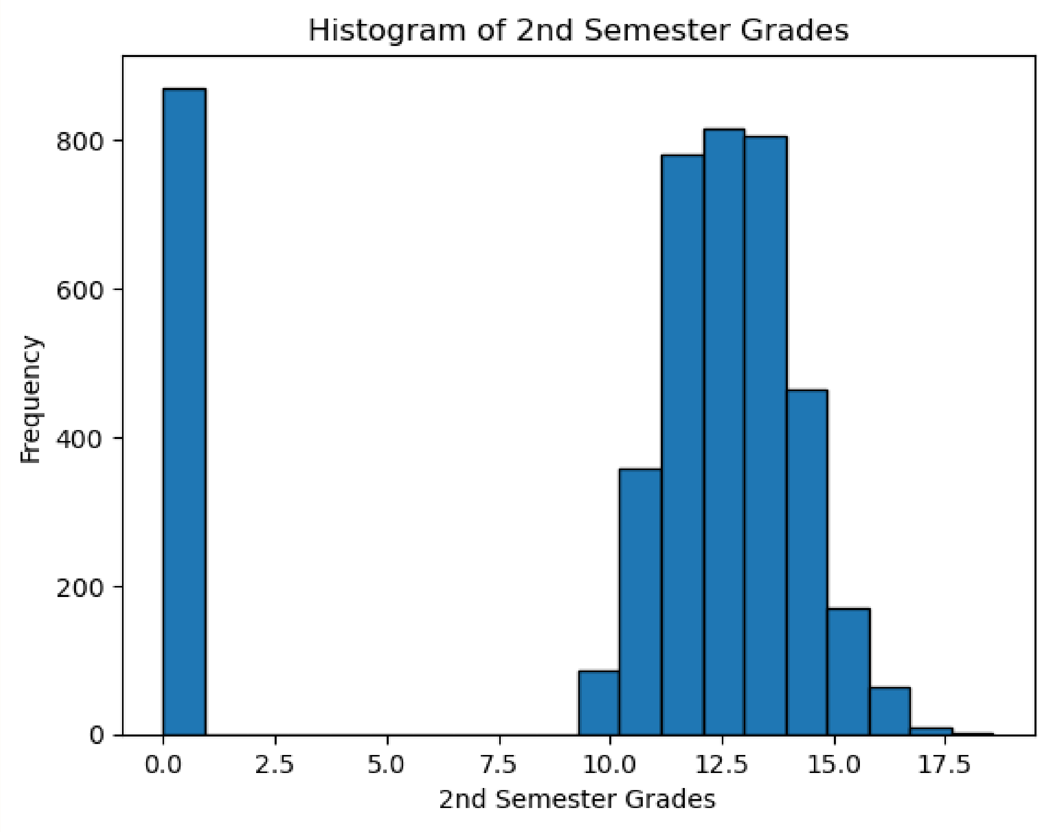
- Categorical, no outliers.
- Mean: 0.25
- Mode: 0
- Standard Deviation: 0.43
- Range: 1
- Skew: 1.16
- Kurtosis: -0.64

CURRICULAR UNITS 1ST SEM (GRADE AVG)



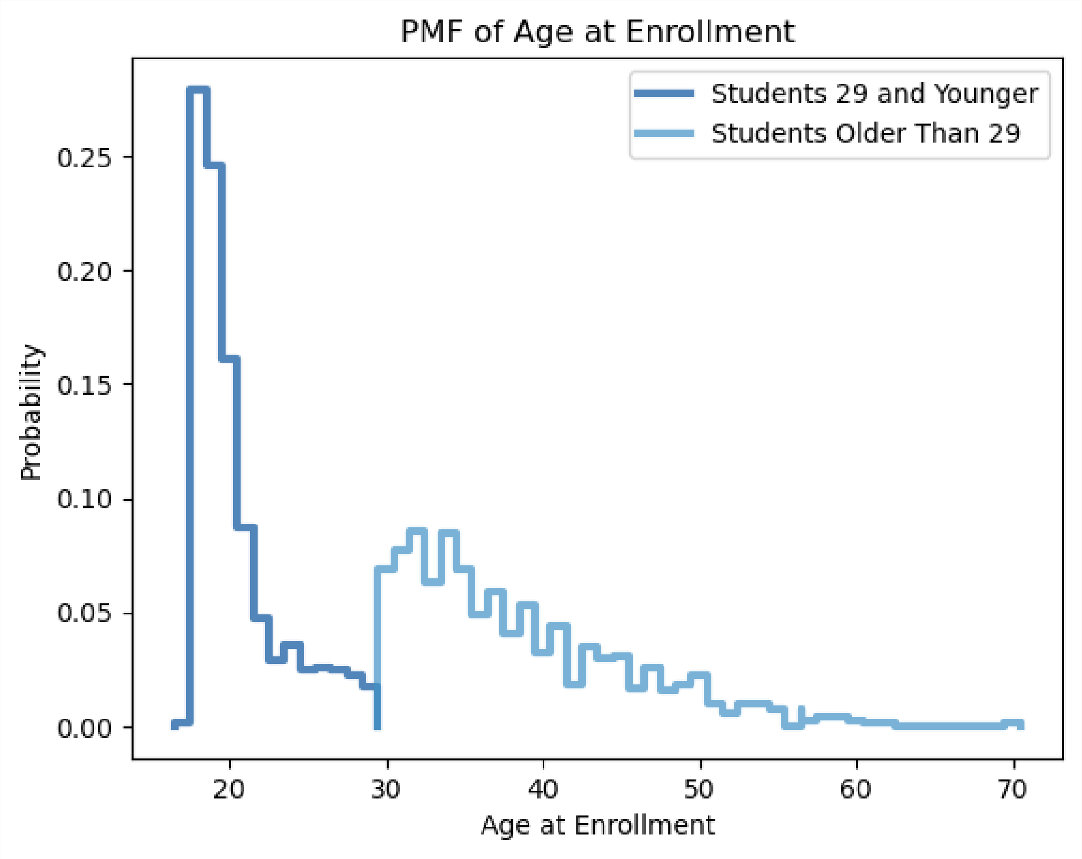
- Outliers: On the right side suggest that there are outliers of students who have scored highly in comparison to those with an avg of 0.
- Mean: 10.64
- Mode: 0
- Standard Deviation: 4.84
- Range: 18.88
- Skew: -1.57
- Kurtosis: 0.91

CURRICULAR UNITS 2ND SEM (GRADE AVG)



- Outliers: On the right side suggest that there are outliers of students who have scored highly in comparison to those with an avg of 0.
- Mean: 10.23
- Mode: 0
- Standard Deviation: 5.21
- Range: 18.57
- Skew: -1.31
- Kurtosis: 0.065

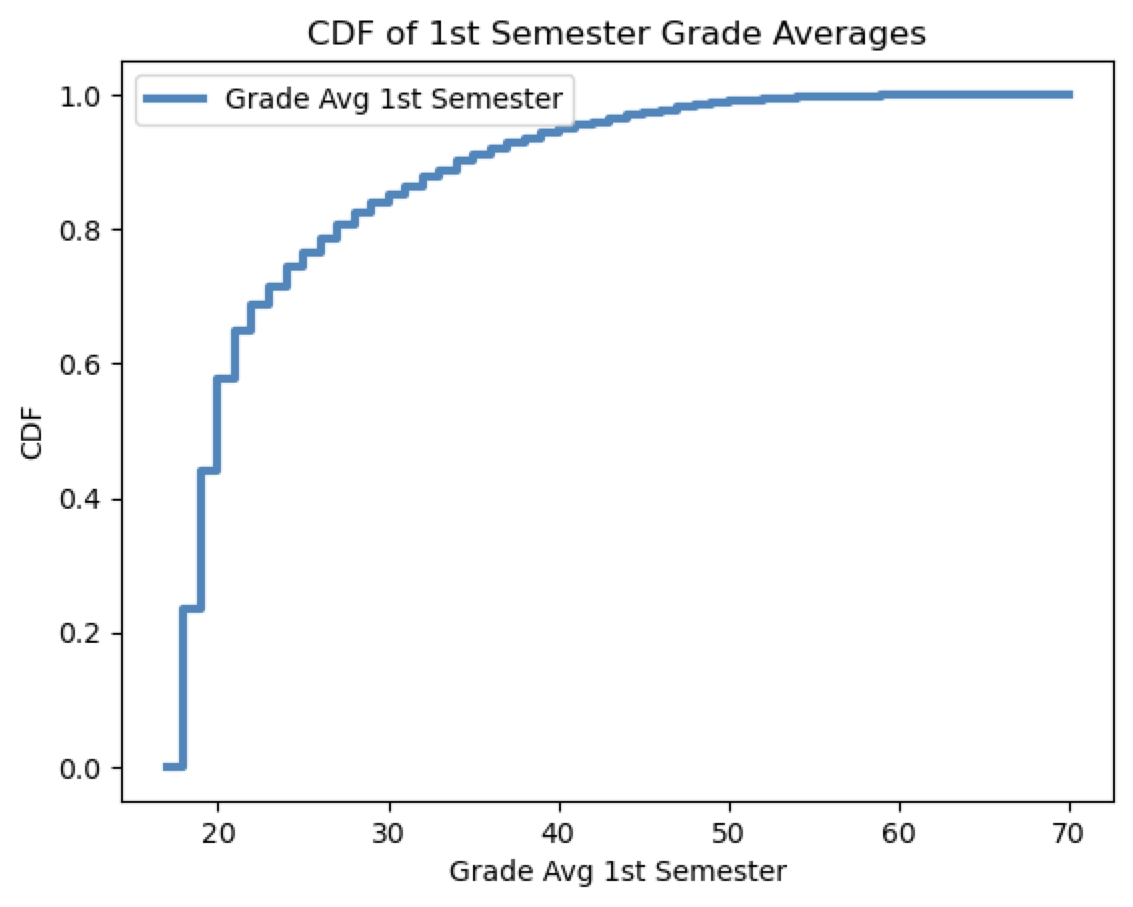
PMF OF AGE AT ENROLLMENT



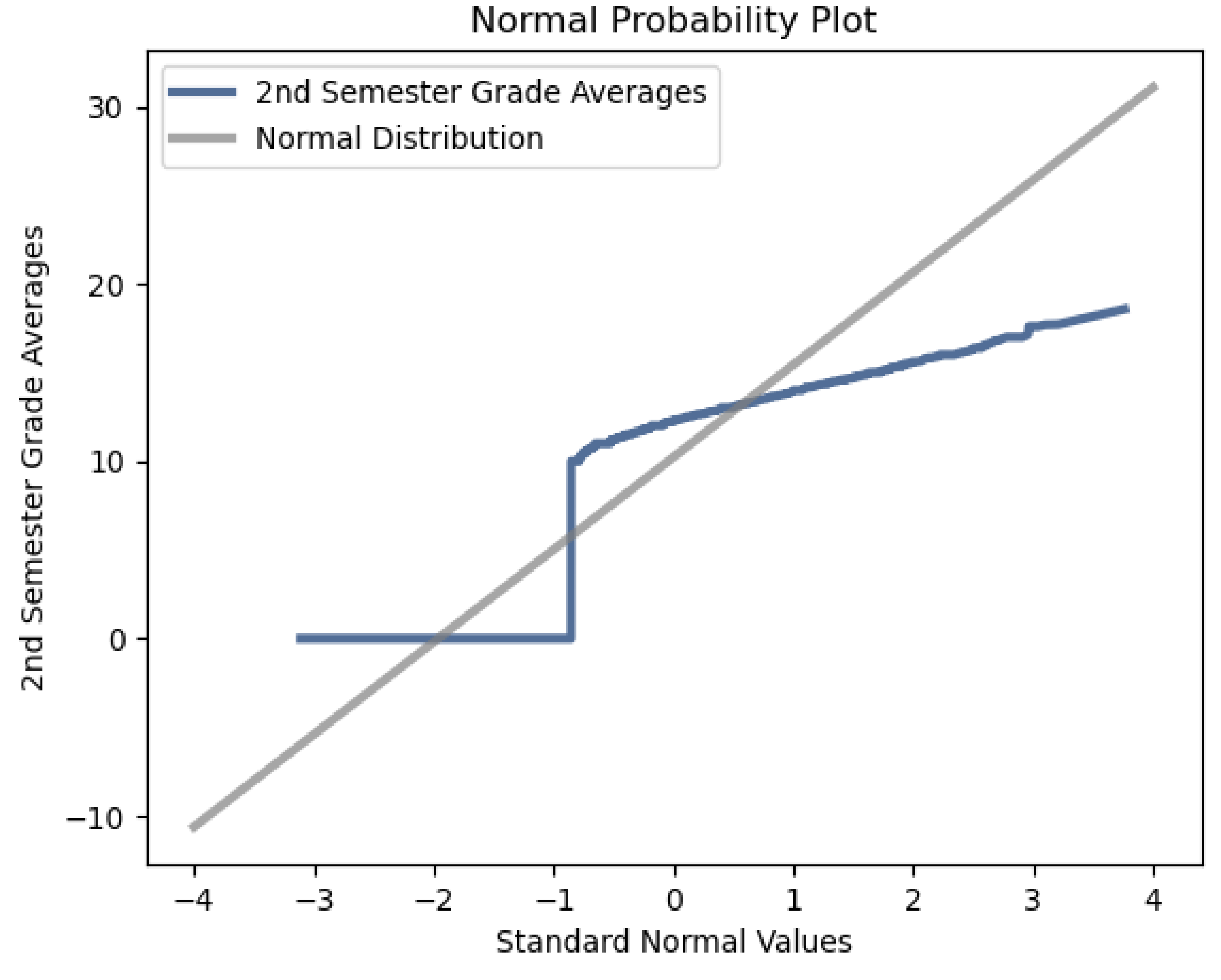
CDF FOR 1ST SEMESTER GRADE AVERAGES

The CDF of 1st Semester Grade Averages indicates that a lot of students scored low because of the steep rise on the left side.

Poor grades in the first semester could attribute to why many students drop out.

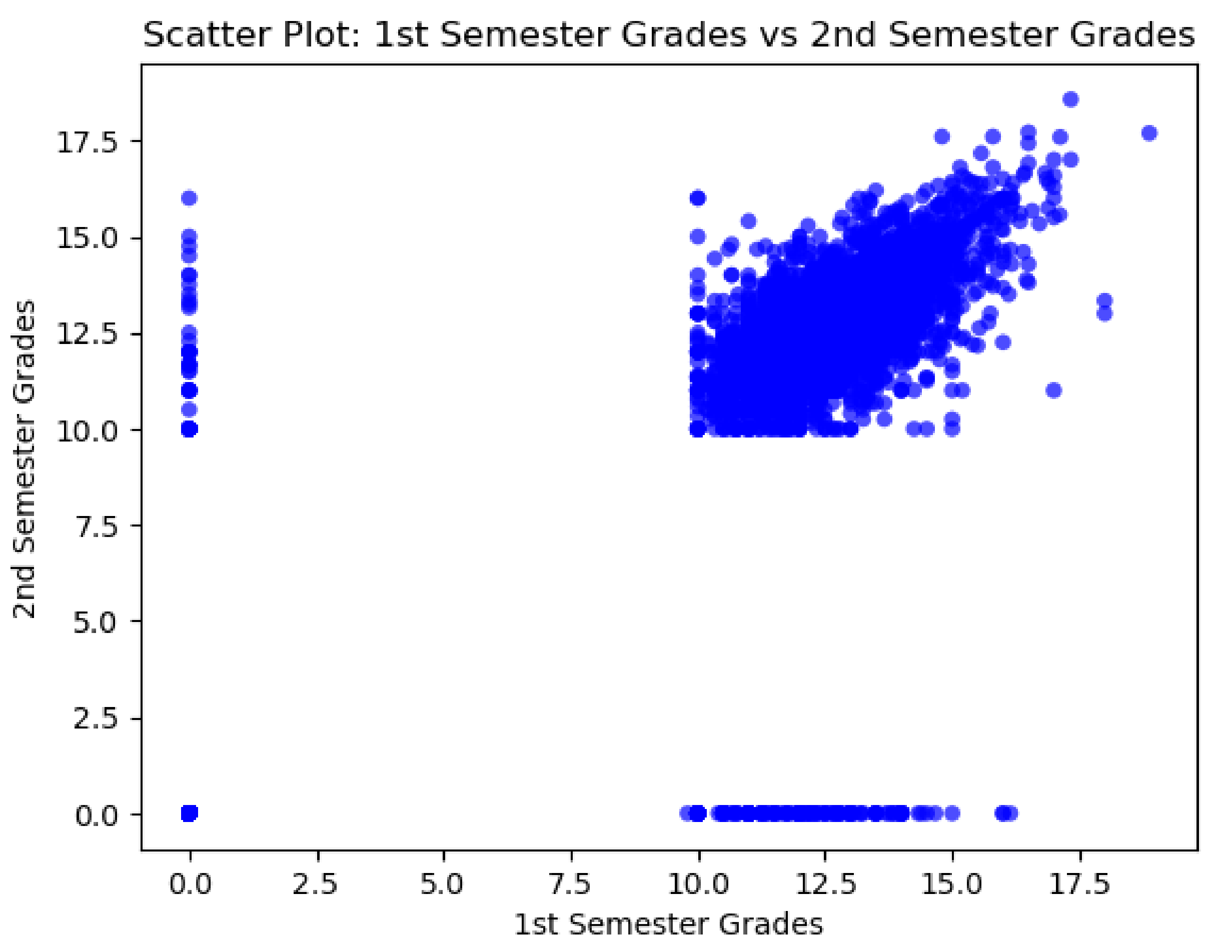


NORMAL DISTRIBUTION OF 2ND SEMESTER GRADE AVERAGES



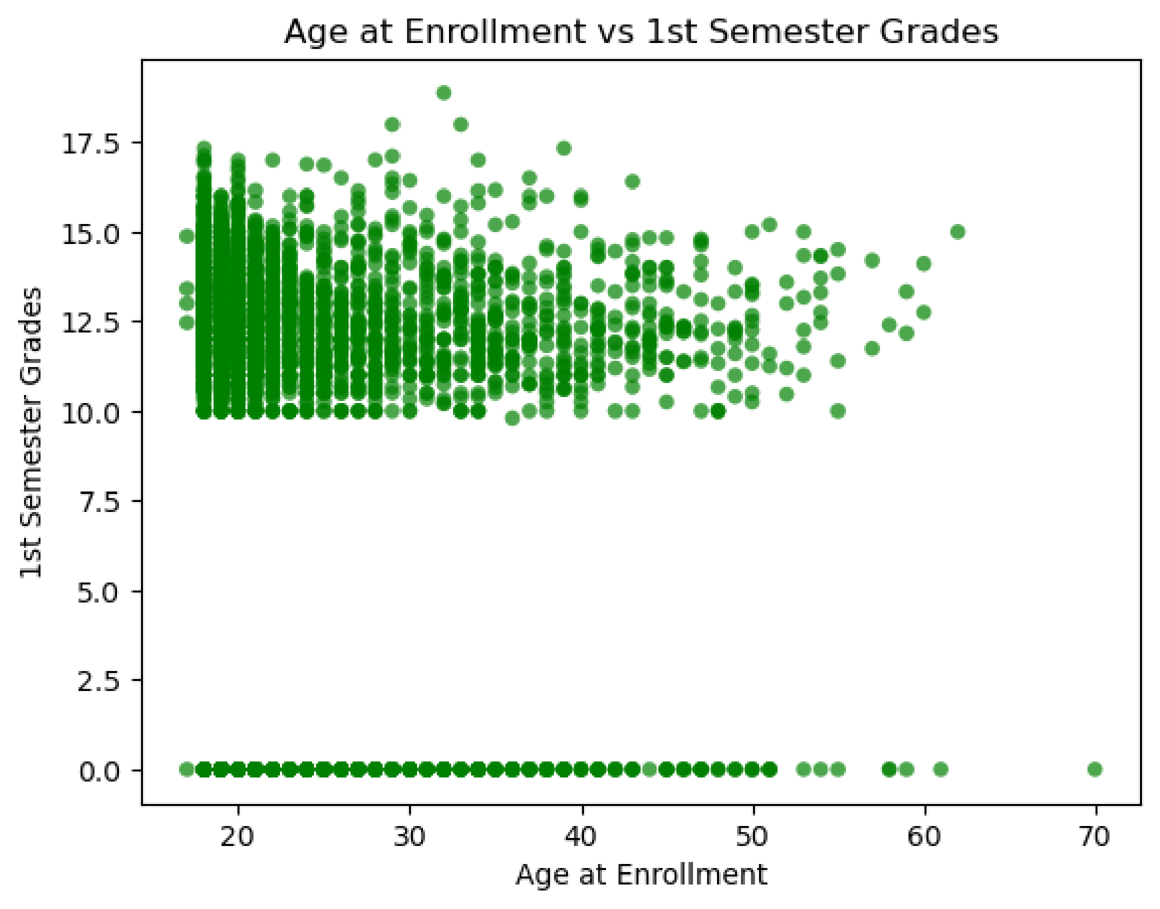
- The gray line represents a normal distribution, while the blue represents the grade averages of the students for 2nd semester.
- Since the blue line doesn't follow the gray much at all, it's clear that the data isn't normally distributed.
- A lot of low grades could contribute to dropout rates.

1ST SEM GRADES VS 2ND SEM



- Correlation: 0.837
- Covariance: 21.129
- A strong positive relationship between the two variables, as shown by correlation coefficient and covariance. Students who tend to perform well in 1st sem also perform well 2nd sem.
- The small dots at the origin indicate students who scored 0 both sems.

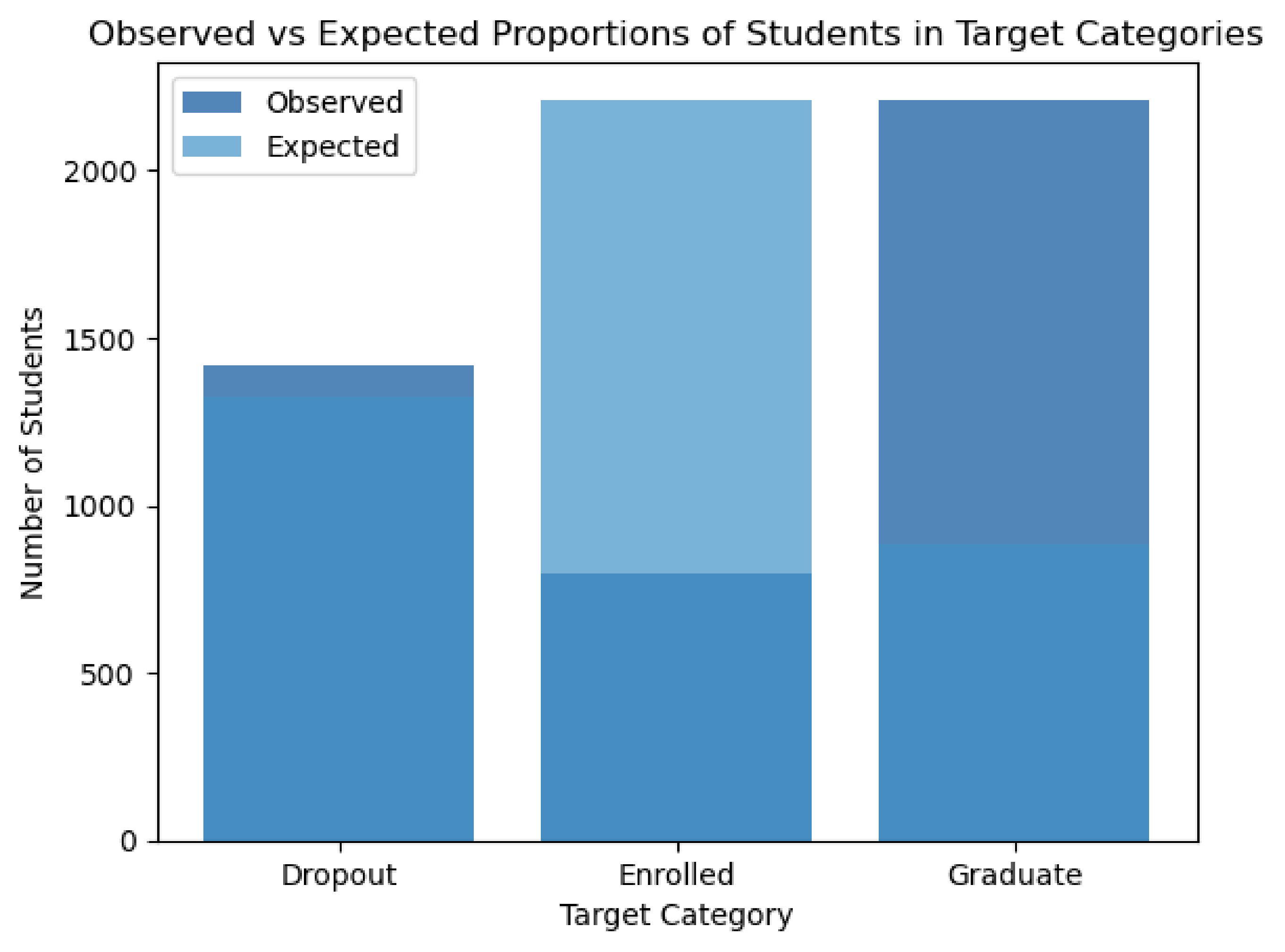
AGE OF ENROLLMENT VS 1ST SEMESTER GRADES



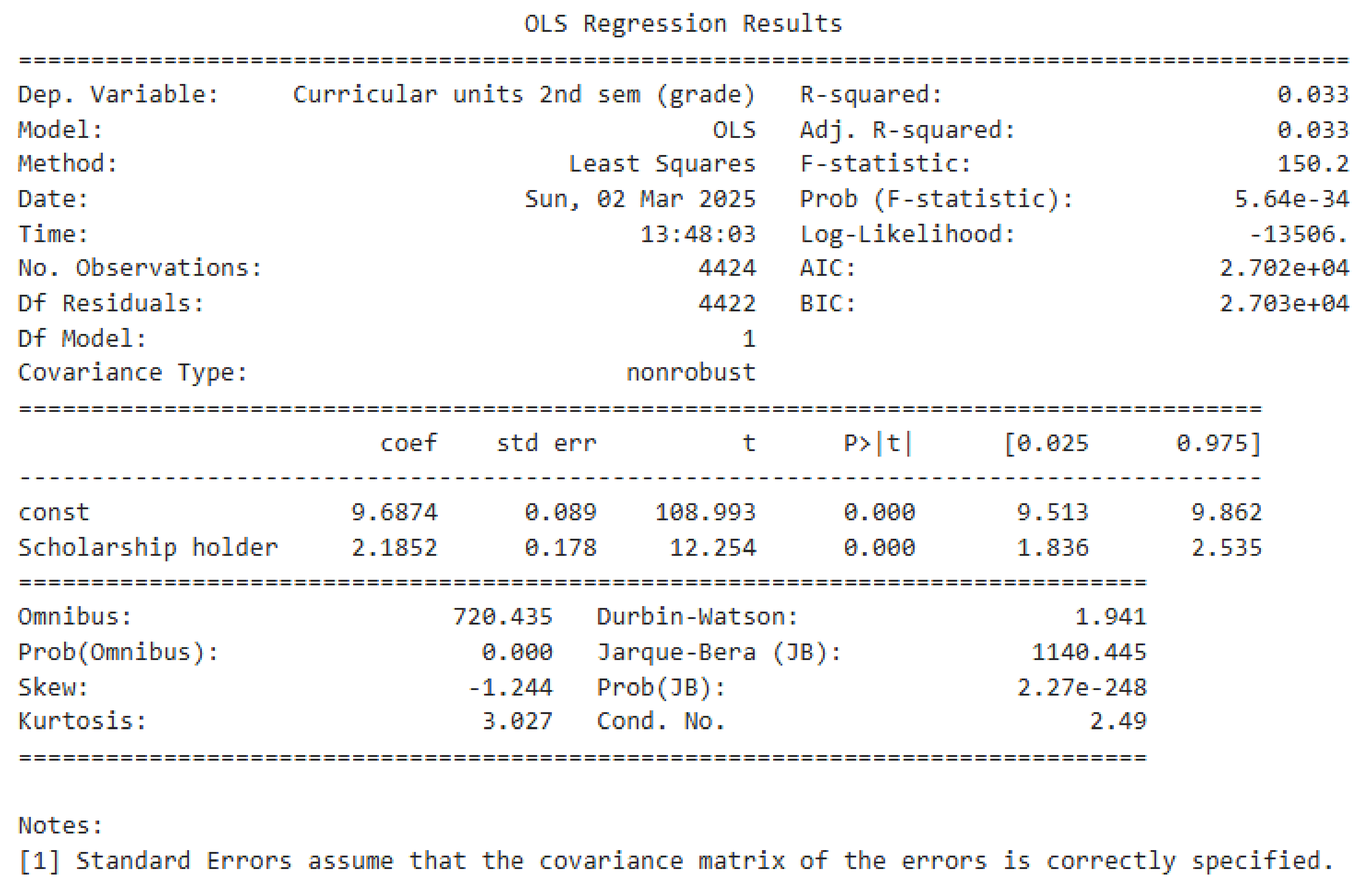
- Correlation: -0.157
- Covariance: -5.756
- There is a weak negative relationship between the two variables.
- Many students, regardless of age have grades around 0.
- Negative covariance means the two variables move in opposite directions, meaning as age increases, grades slightly decrease.

HYPOTHESIS TEST

- Used Chi-Squared Test to test the proportion of the variable Target (Dropout, Enrolled, Graduate)
- The Chi-Square Statistic: 2897.447
- P-Value: 0.0
- The observed proportions are significantly different than the expected proportions.
- The distribution of students across the categories don't follow the expected distribution.



REGRESSION ANALYSIS: 2ND SEMESTER GRADES AND SCHOLARSHIP HOLDERS



```
                                        OLS Regression Results
=============================================================================================
Dep. Variable:     Curricular units 2nd sem (grade)   R-squared:                        0.033
Model:                                          OLS   Adj. R-squared:                   0.033
Method:                               Least Squares   F-statistic:                      150.2
Date:                              Sun, 02 Mar 2025   Prob (F-statistic):            5.64e-34
Time:                                      13:48:03   Log-Likelihood:                 -13506.
No. Observations:                              4424   AIC:                          2.702e+04
Df Residuals:                                  4422   BIC:                          2.703e+04
Df Model:                                         1
Covariance Type:                          nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  9.6874      0.089    108.993      0.000       9.513       9.862
Scholarship holder     2.1852      0.178     12.254      0.000       1.836       2.535
==============================================================================
Omnibus:                      720.435   Durbin-Watson:                   1.941
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1140.445
Skew:                          -1.244   Prob(JB):                    2.27e-248
Kurtosis:                       3.027   Cond. No.                         2.49
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```