OESON Learning Data Science:
Student Enrollment, Exploratory Data Analysis, and Statistical Modeling

Performed by Thomas Saraceno

Table of Contents

Table of Contents (continued)

Prompt

- A dataset is given with various attributes related to students enrolled at a higher education institution, including:
  - Demographic: Marital Status, Nationality, Displaced Status, Gender, Age at Enrollment, and International Status
  - Socioeconomic: Mother's Qualification, Father's Qualification, Mother's Occupation, Father's Occupation, Educational Special Needs Status, Debtor Status, Tuition Fees Up to Date Status, and Scholarship Holder Status
  - Initial Enrollment: Application Mode, Application Order, Course, Daytime/Evening Attendance, Previous Qualification
  - Macroeconomic: Unemployment Rate, Inflation Rate, GDP
  - First and Second Semester: Curricular Units Credited, Enrolled Approved, Evaluated, Not Evaluated, and Grade

Prompt (continued)

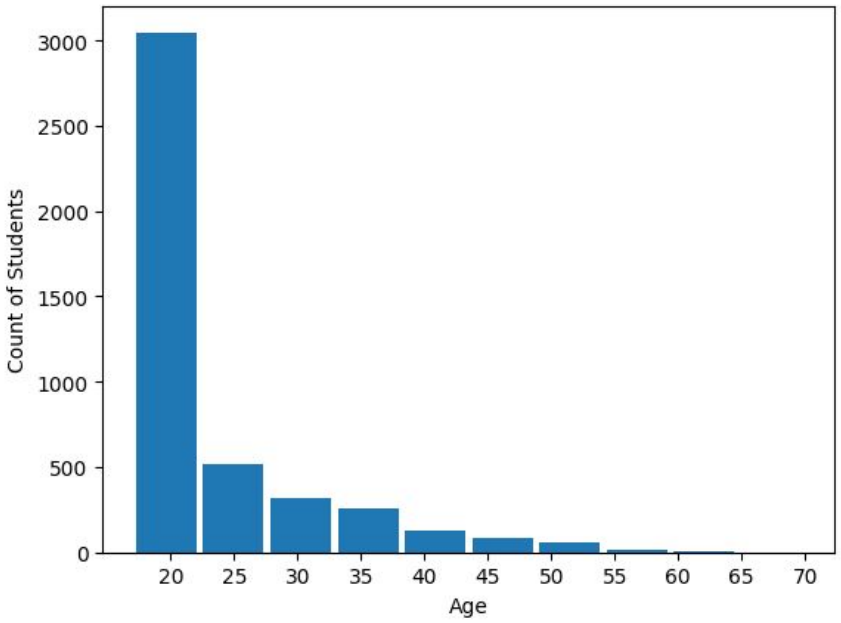
- The dataset aims to analyze and understand factor influencing student enrollment, academic performance, and socio-economic background in higher education institutions
- The goal is to identify patterns, trends, and potential areas for intervention or improvement in the educational system
- The dataset can be utilized to develop predictive models to forecast enrollment patterns, student success rates, and the impact of economic indicators on the education sector
- The analysis can provide valuable insights for educational policymakers, administrators, and stakeholders to enhance educational access, equity, and quality

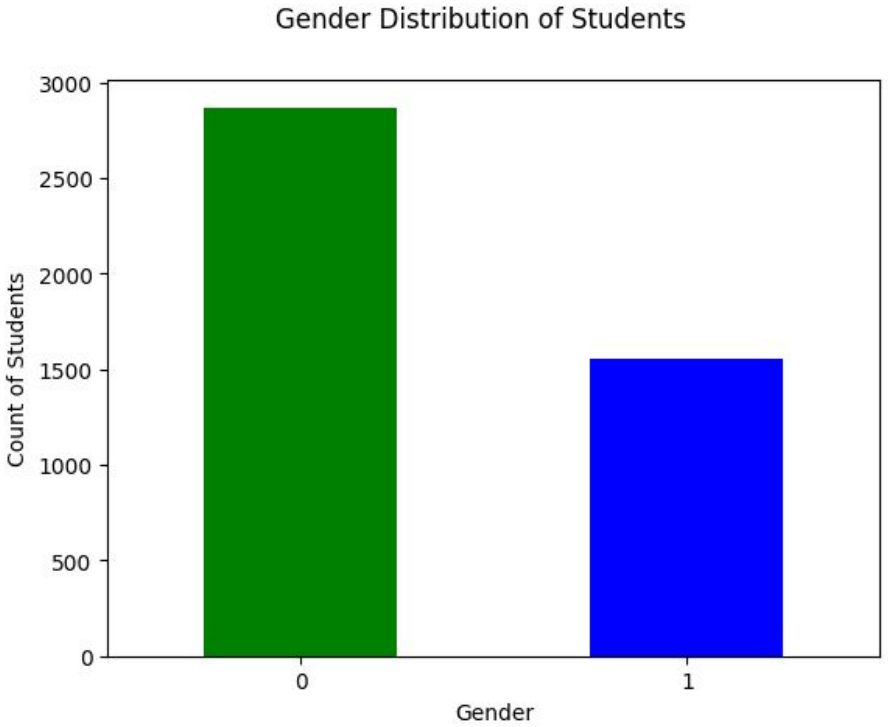
Demographic Data EDA

- To begin, the basic characteristics and statistics of the dataset will be shown through charts and graphs
- This is called Exploratory Data Analysis (EDA)

Descriptor	Mean	Median	Mode	Minimum	Maximum
Marital Status	1.19	1.0	1	1	1
Nationality	1.25	1.0	1	1	21
Displaced	0.54	1.0	1	0	1
Gender	0.35	0.0	0	0	1
Age at Enrollment	23.27	20.0	18	17	70
International	0.02	0.0	0	0	1

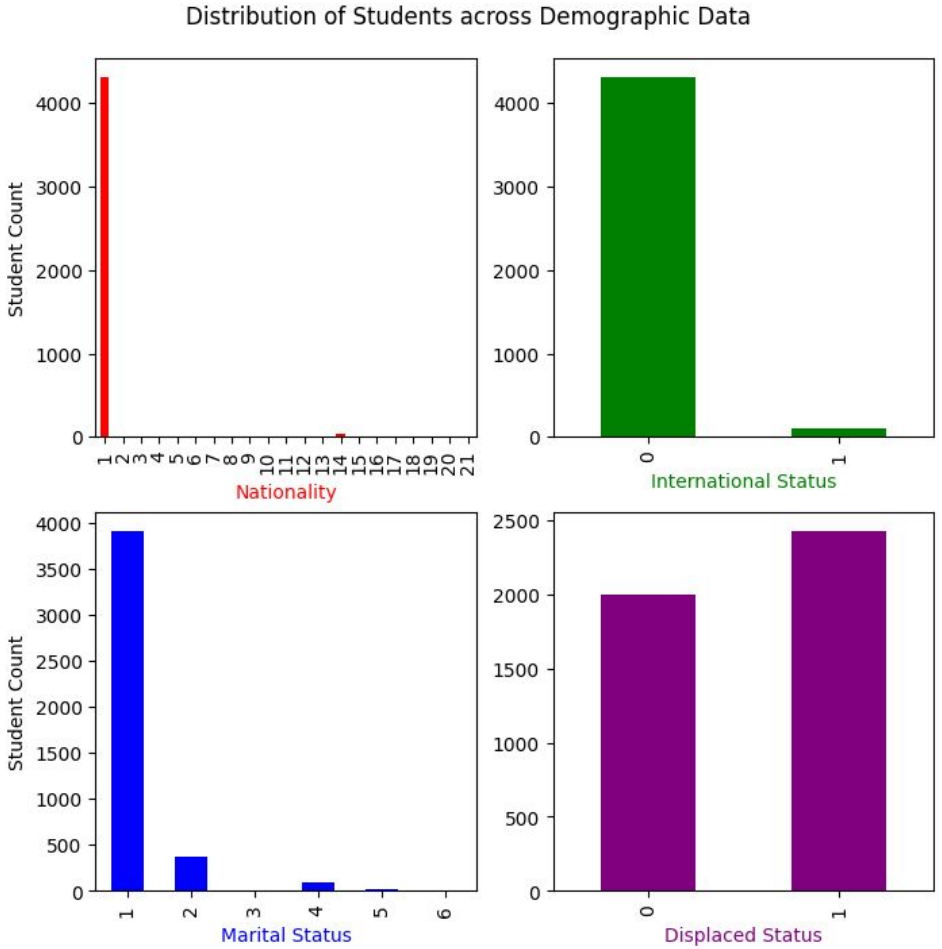
Demographic Data EDA (continued)





Demographic Data EDA (continued)

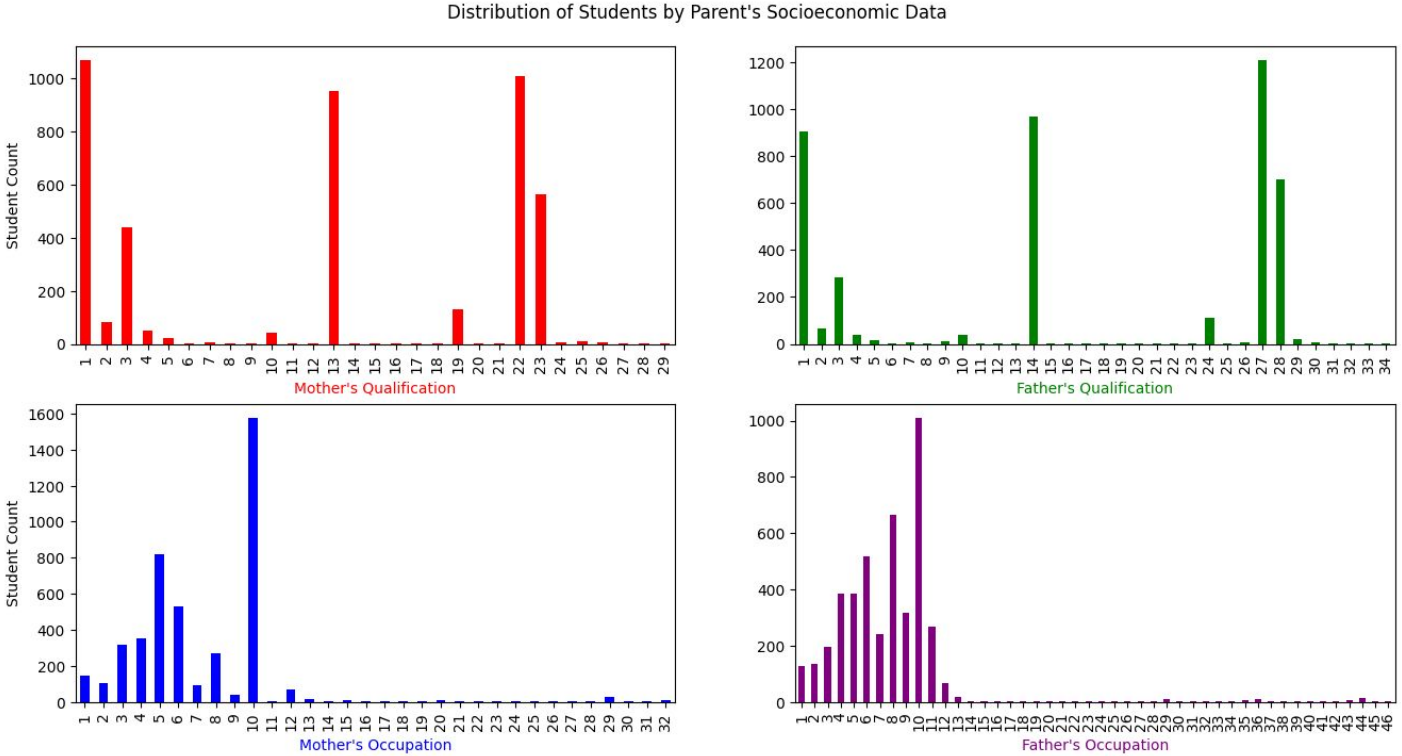
- From these statistics we infer:
  - Students tend to be younger
  - Gender 0 is more represented in the data
  - Most are from Nationality 1
  - Most have International Status 0
  - Most have Marital Status 1
  - There is a slightly higher amount of students with Displaced Status 1



Socioeconomic Data EDA

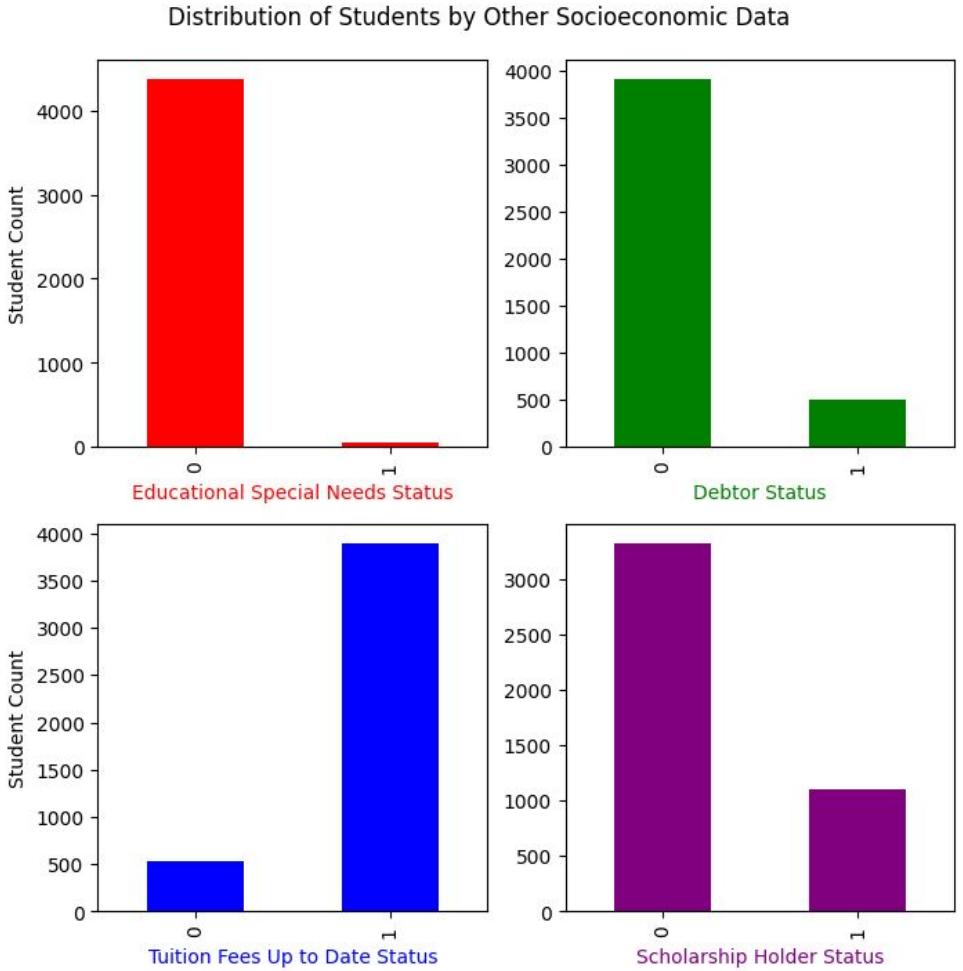
Descriptor	Mean	Median	Mode	Minimum	Maximum
Mother's Qualification	12.32	13.0	1	1	29
Father's Qualification	16.46	14.0	27	1	34
Mother's Occupation	7.32	6.0	10	1	32
Father's Occupation	7.82	8.0	10	1	46
Educational Special Needs	0.01	0.0	0	0	1
Debtor	0.11	0.0	0	0	1
Tuition Fees Up to Date	0.88	1.0	1	0	1
Scholarship Holder	0.25	0.0	0	0	1

Socioeconomic Data EDA (continued)



Socioeconomic Data EDA (continued)

- From these statistics we infer:
  - Mother's Qualification 1, 13, and 22 are most frequent
  - Father's Qualification 1, 14, and 27 are most frequent
  - Mother's Occupation 10 is most frequent
  - Father's Occupation 10 is most frequent
  - Most students are of Educational Special Needs Status 0
  - Most students are of Debtor Status 0
  - Most students are of Tuition Fees Up to Date Status 1
  - There are more students with Scholarship Holder Status 0

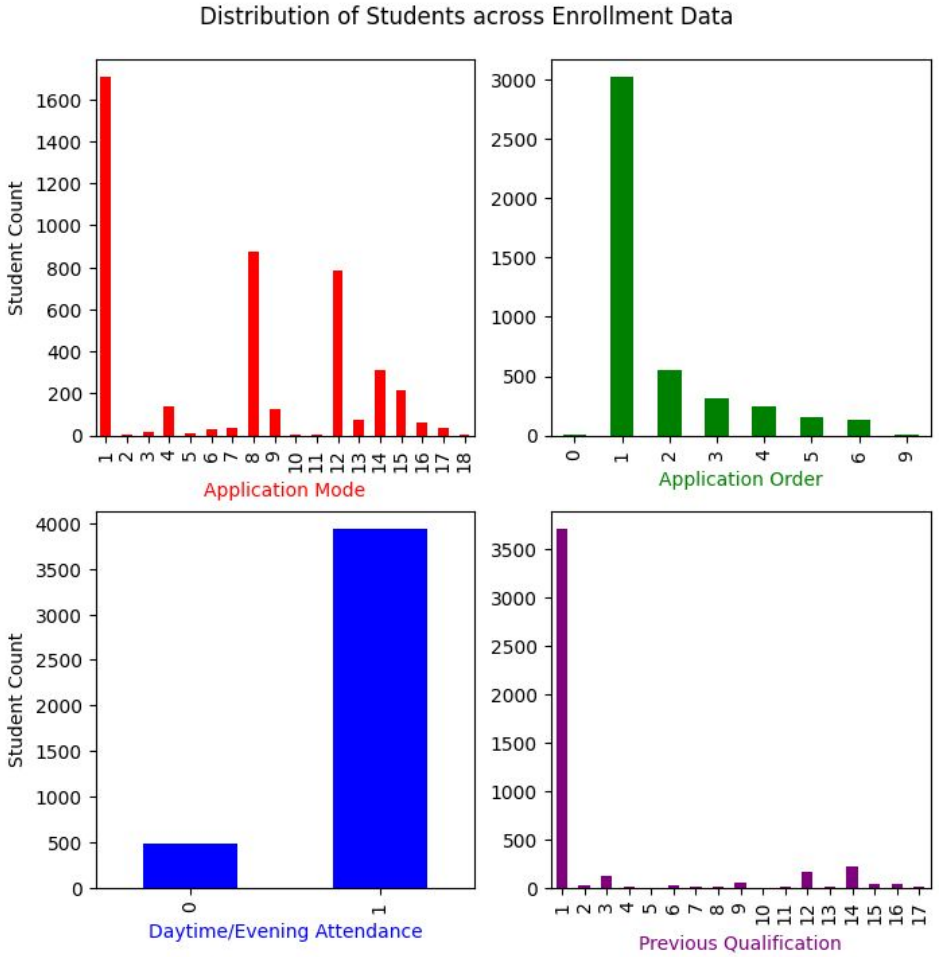


Initial Enrollment EDA

Descriptor	Mean	Median	Mode	Minimum	Maximum
Application Mode	6.87	8.0	1	1	18
Application Order	1.73	1.0	1	0	9
Course	9.90	10.0	12	1	17
Daytime/Evening Attendance	0.89	1.0	1	0	1
Previous Qualification	2.53	1.0	1	1	17

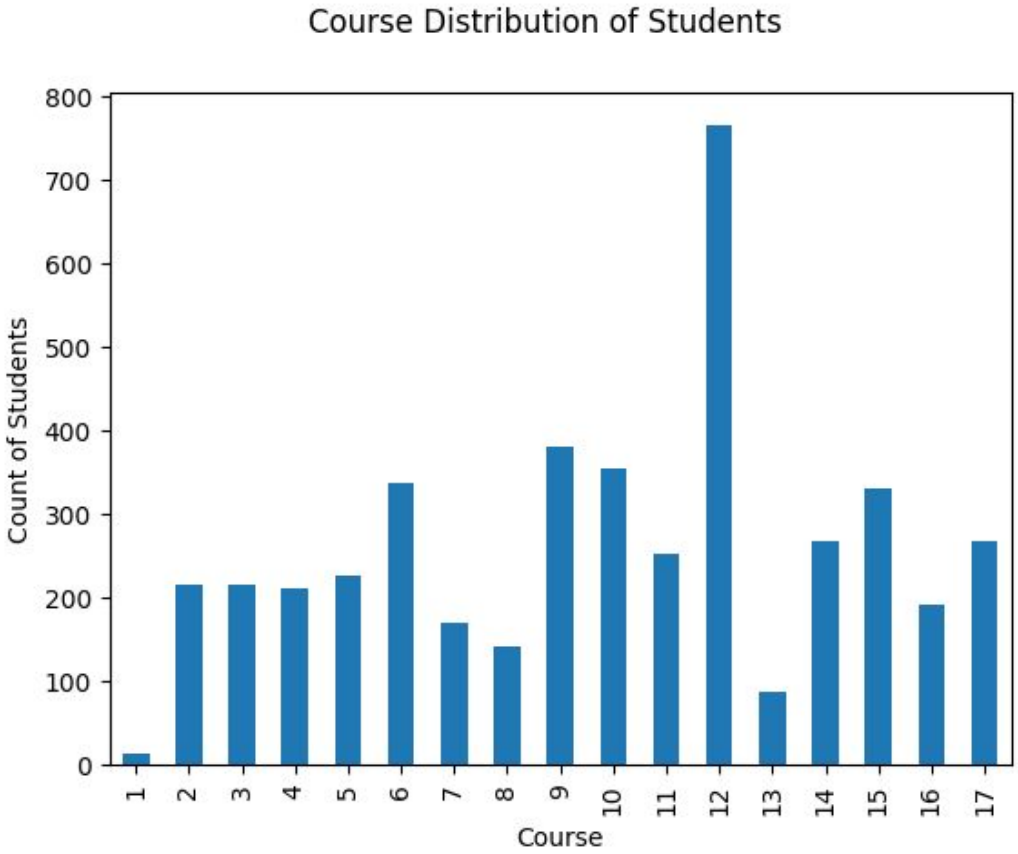
Initial Enrollment EDA (continued)

- From these statistics we infer:
  - Application mode 1, 8, and 12 occur most frequently
  - Application order 1 is most frequent
  - Most students are of Daytime/Evening Attendance Status 1
  - Most students are of Previous Qualification number 1



Initial Enrollment EDA (continued)

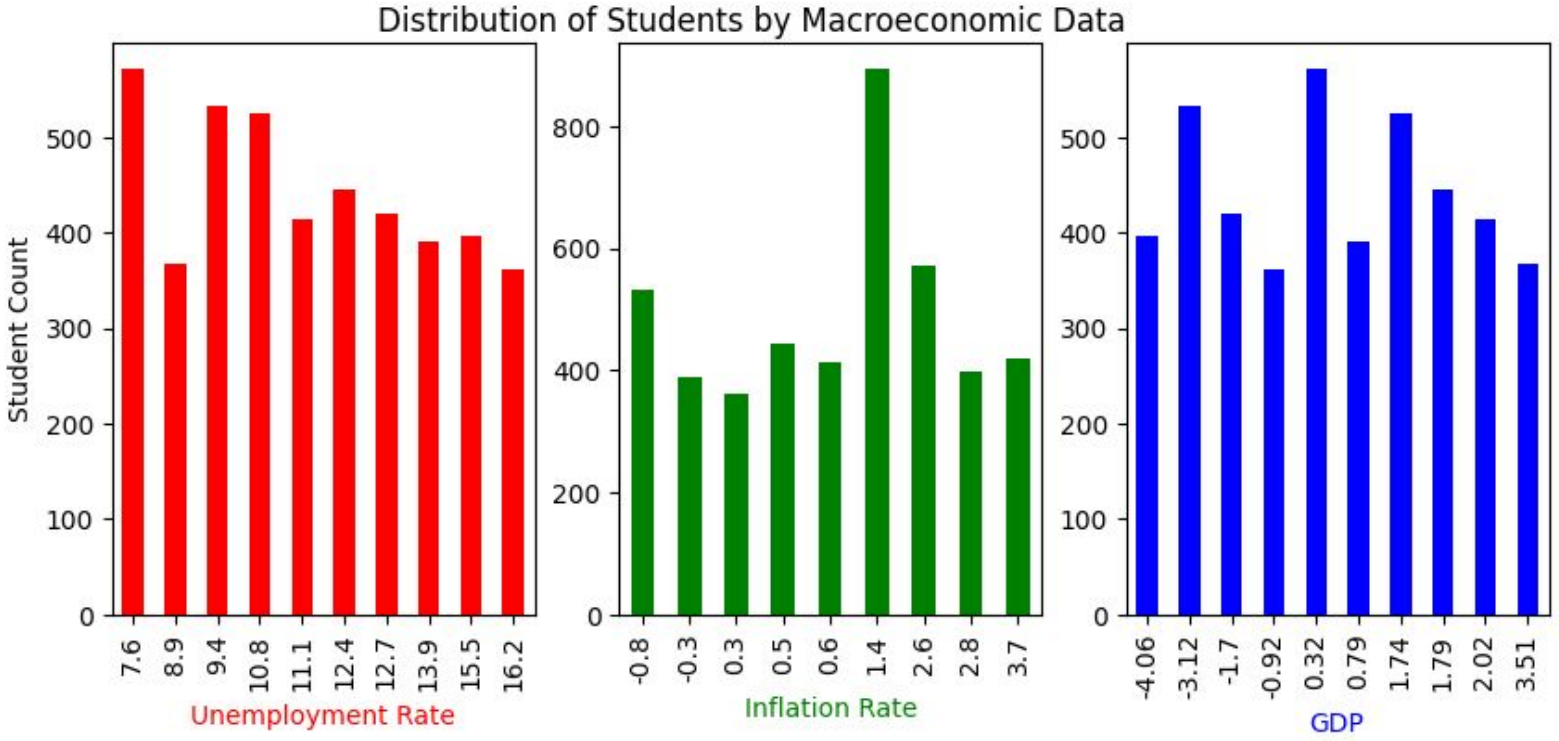
- We also infer:
  - Course number 12 has the highest number of students enrolled
  - Course number 1 has the lowest number



Macroeconomic EDA

Descriptor	Mean	Median	Mode	Minimum	Maximum
Unemployment Rate	11.57	11.1	7.6	7.6	16.2
Inflation Rate	1.23	1.4	1.4	-0.8	3.7
GDP	0.002	0.32	0.32	-4.06	3.51

Macroeconomic EDA (continued)

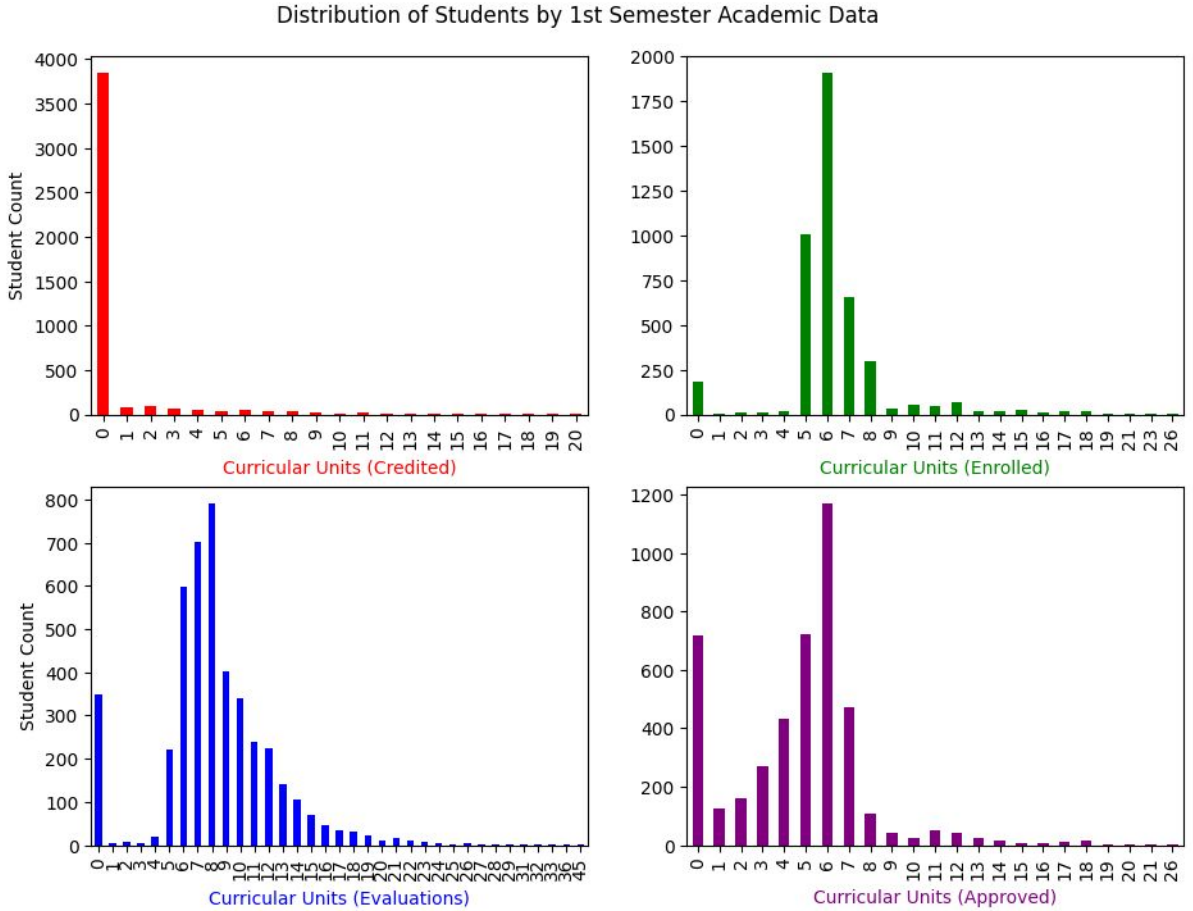


First Semester EDA

Descriptor	Mean	Median	Mode	Minimum	Maximum
Curricular Units 1st Sem (Credited)	0.71	0.0	0	0	20
Curricular Units 1st Sem (Enrolled)	6.27	6.0	6	0	26
Curricular Units 1st Sem (Evaluations)	8.30	8.0	8	0	45
Curricular Units 1st Sem (Approved)	4.71	5.0	6	0	26
Curricular Units 1st Sem (Grade)	10.64	12.29	0.0	0.0	18.88
Curricular Units 1st Sem (w/o Evaluations)	0.14	0.0	0	0	12

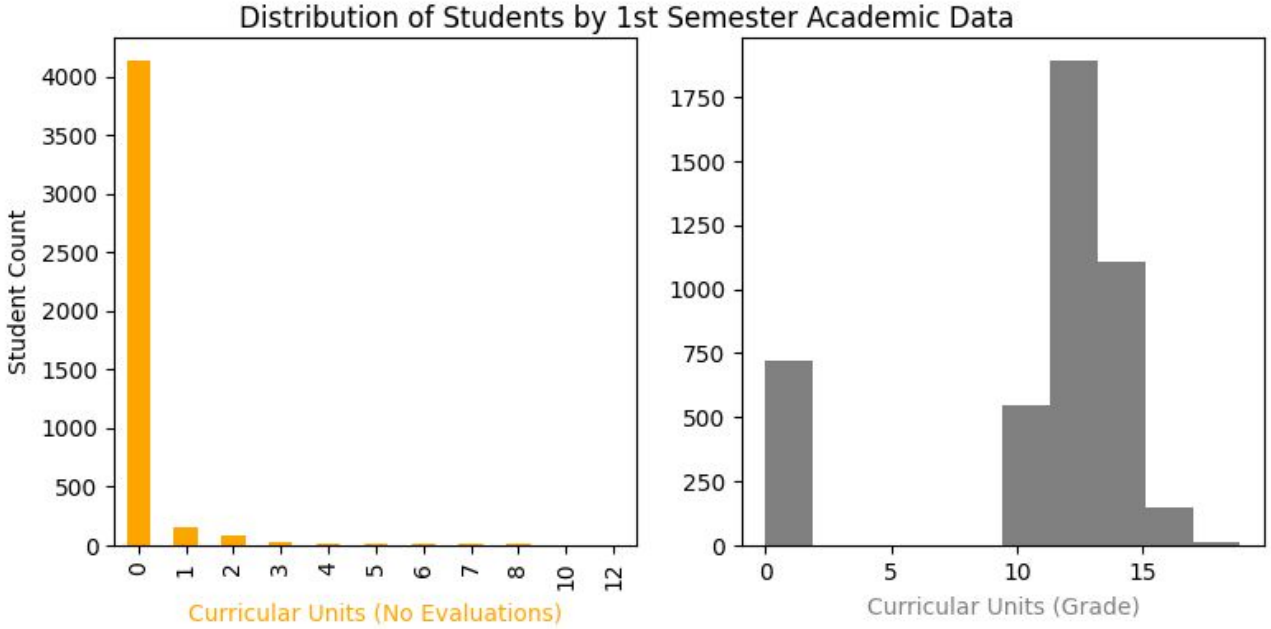
First Semester EDA (continued)

- From these statistics we infer:
  - The most frequent number of curricular units credited is 0
  - The most frequent number of curricular units enrolled is 6
  - The most frequent numbers of curricular units evaluated are 6, 7, and 8
  - The most frequent numbers of curricular units approved are 0, 5, and 6



First Semester EDA (continued)

- We also infer:
  - The most frequent number of curricular units without evaluations is 0
  - Grades are most frequently occuring in the 10-15 mark range, with a not insignificant number in the 0 range

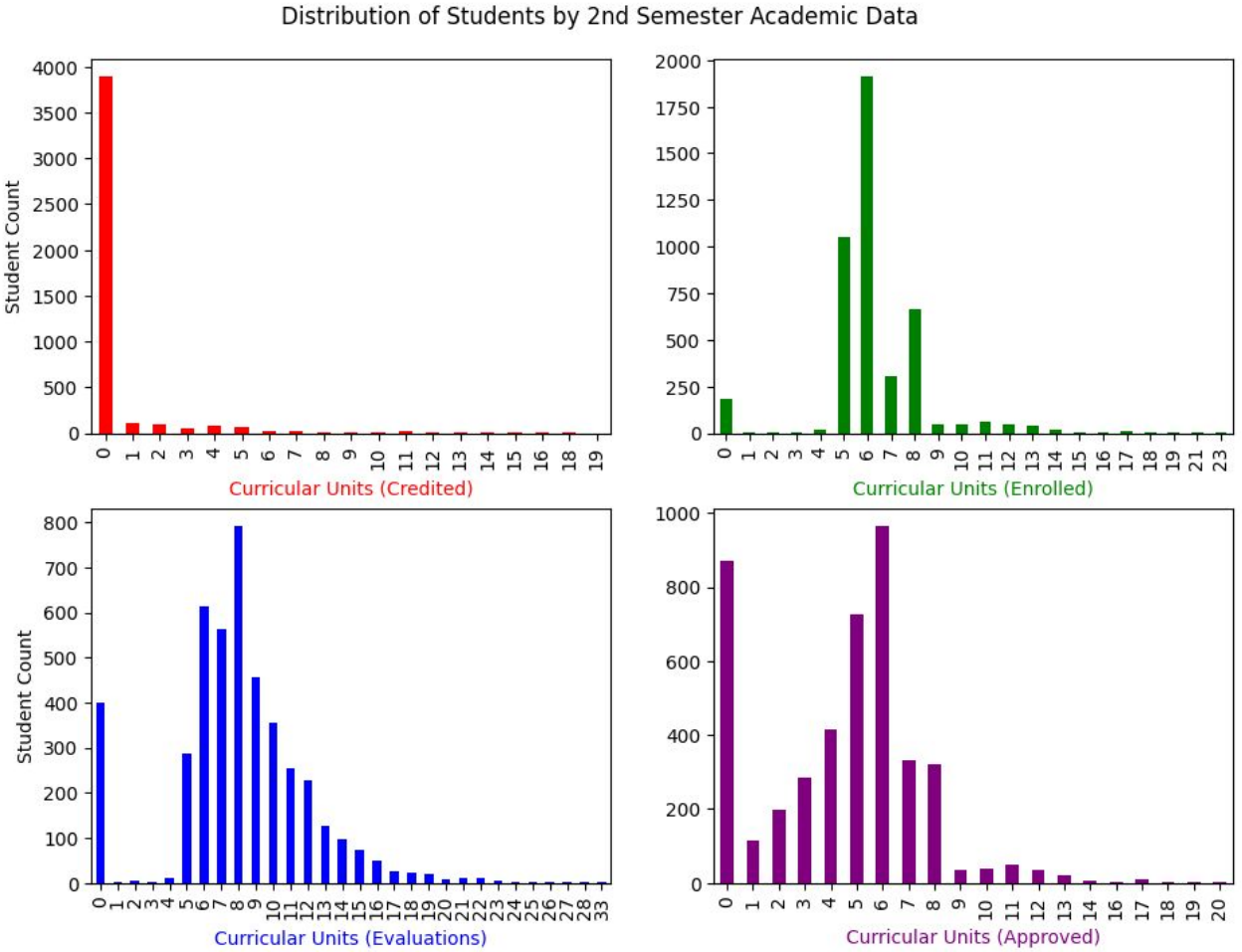


Second Semester EDA

Descriptor	Mean	Median	Mode	Minimum	Maximum
Curricular Units 2nd Sem (Credited)	0.54	0.0	0	0	19
Curricular Units 2nd Sem (Enrolled)	6.23	6.0	6	0	23
Curricular Units 2nd Sem (Evaluations)	8.06	8.0	8	0	33
Curricular Units 2nd Sem (Approved)	4.44	5.0	6	0	20
Curricular Units 2nd Sem (Grade)	10.23	12.2	0.0	0.0	18.57
Curricular Units 2nd Sem (w/o Evaluations)	0.15	0.0	0	0	12

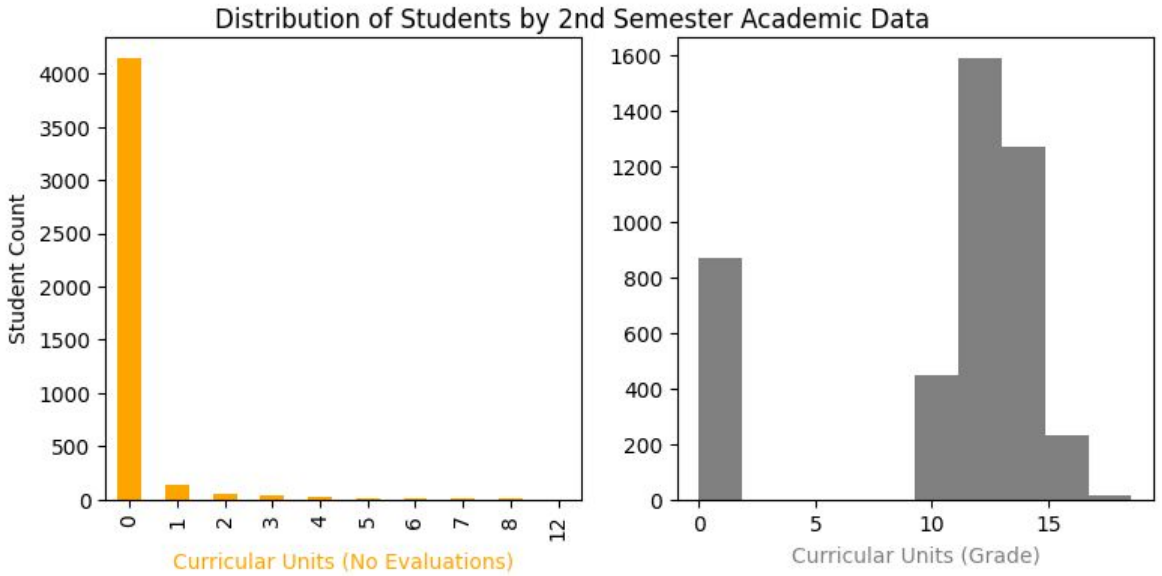
Second Semester EDA (continued)

- From these statistics we infer:
  - The most frequent number of curricular units credited is 0
  - The most frequent number of curricular units enrolled is 6
  - The most frequent numbers of curricular units evaluated are 6, 7, and 8
  - The most frequent numbers of curricular units approved are 0, 5, and 6



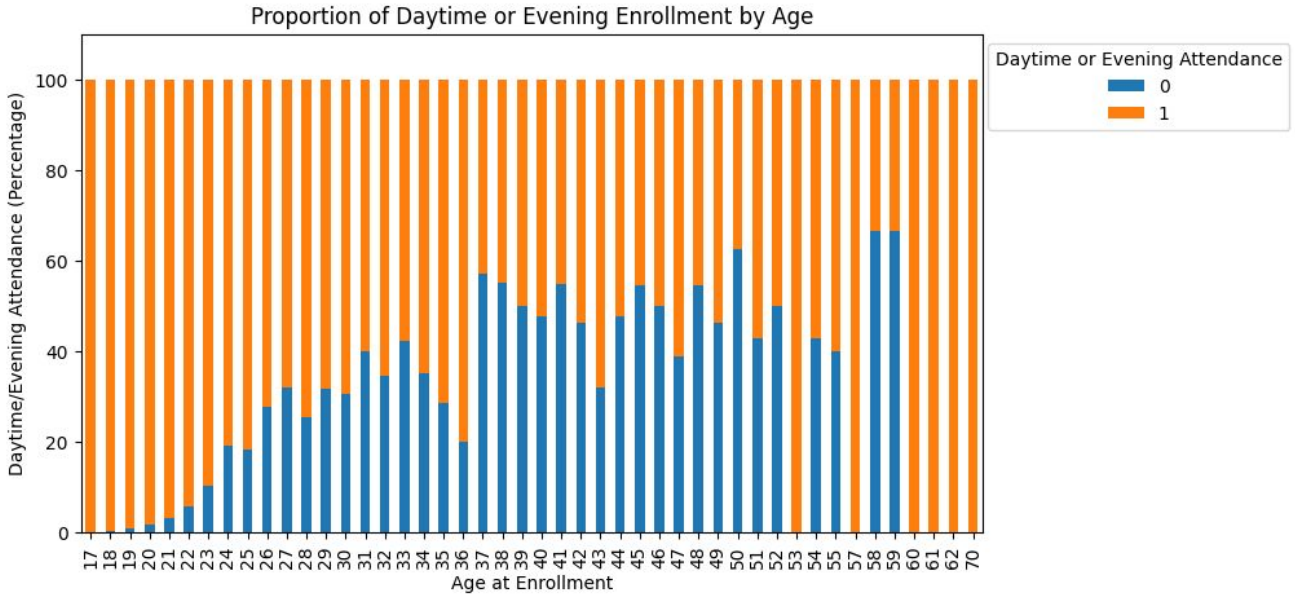
Second Semester EDA (continued)

- We also infer:
  - The most frequent number of curricular units without evaluations is 0
  - Grades are most frequently occuring in the 10-15 mark range, with a not insignificant number in the 0 range



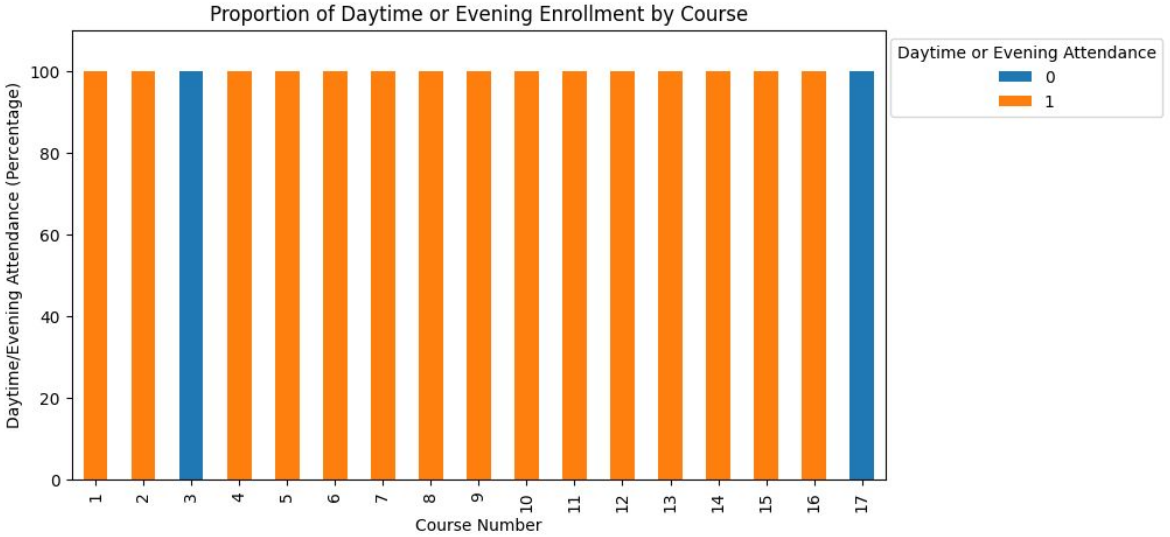
Daytime/Evening Enrollment by Age

- From these statistics we infer:
  - There is a higher proportion of younger-aged students in the 1 category of daytime/evening attendance
  - There is marked increase in the proportion of students in the 0 category of daytime/evening attendance as student age increases



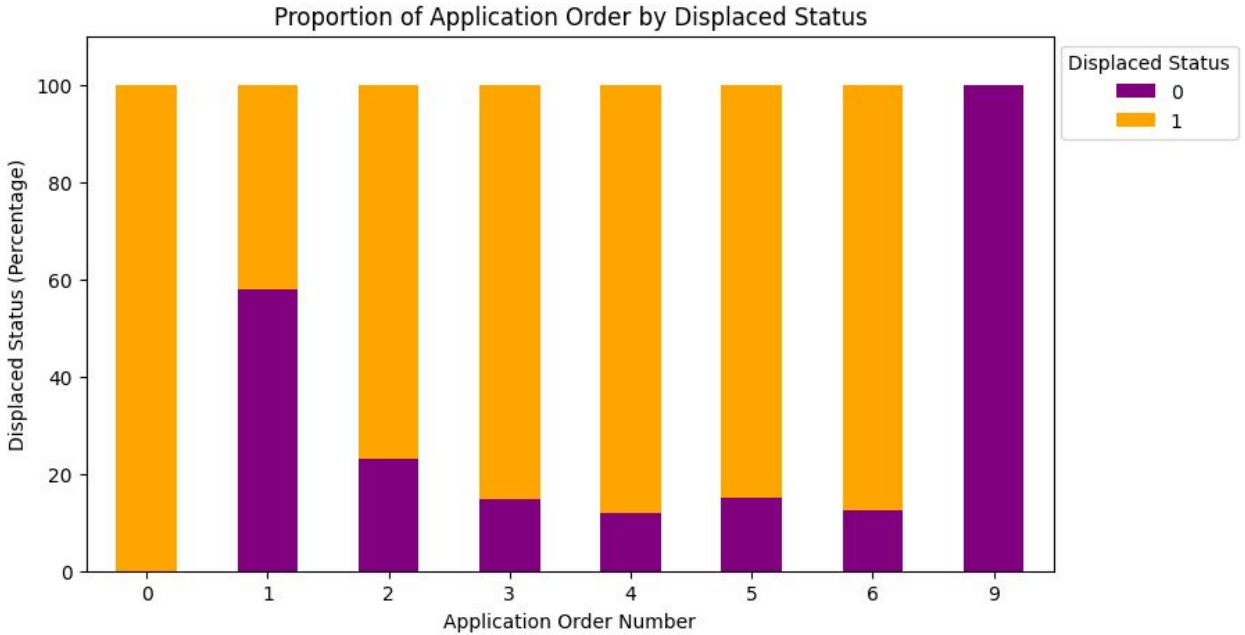
Daytime/Evening Enrollment by Course

- We also infer:
  - Course numbers 3 and 17 are taken by category 0 daytime/evening attendance students, whereas the rest are taken by category 1, showing a split



Application Order by Displaced Status

- From these statistics we infer:
  - Application order 9 is used most frequently by students of displaced status 0, with order 1 and 2 following
  - Category 0 is used most frequently by students of status 1, with 4 and 6 following

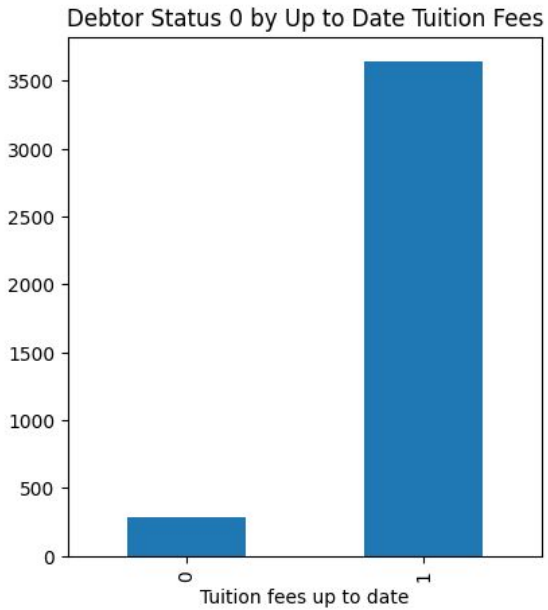


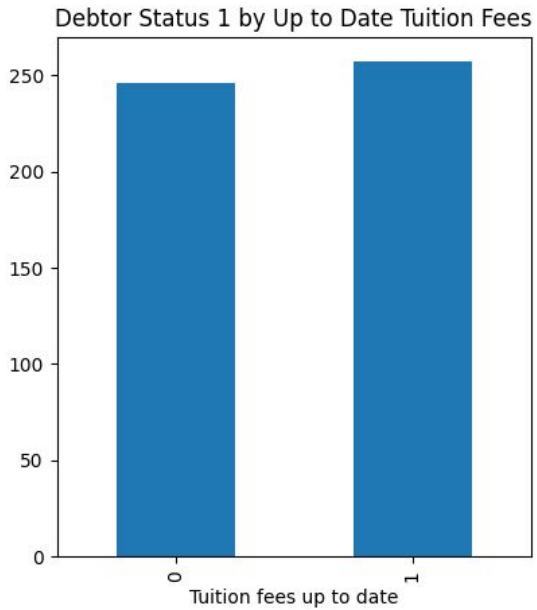
Unemployment Rate by Displaced Status

- From these statistics we infer:
  - There is an observable trend showing that as unemployment rate increases, so does the proportion of students with displaced status 0

Debtor Status by Up to Date Tuition Fees Status

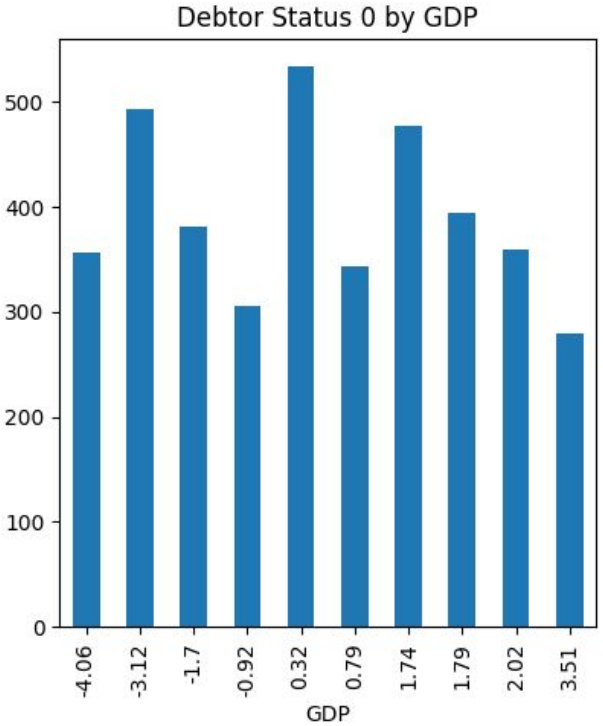
- From these statistics we infer:
  - Those of debtor status 0 tend to be of tuition fees up to date status 1
  - Those of debtor status 1 have a more even split among tuition fees up to date status

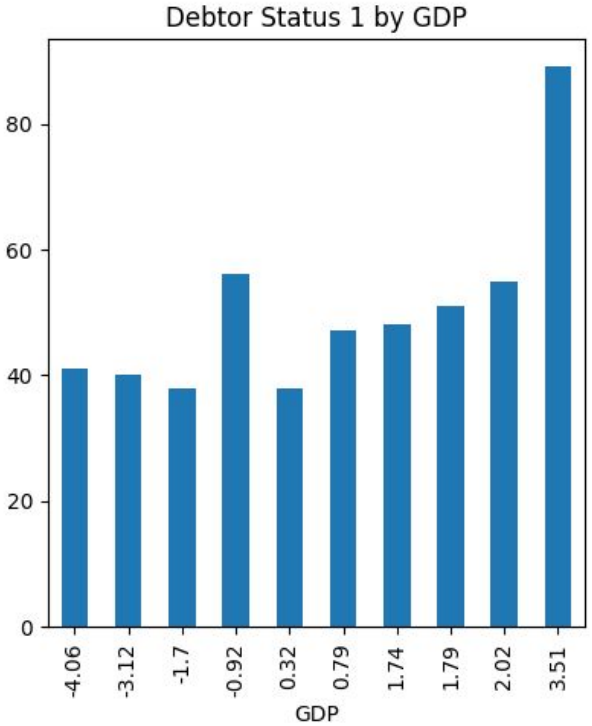




Debtor Status by GDP

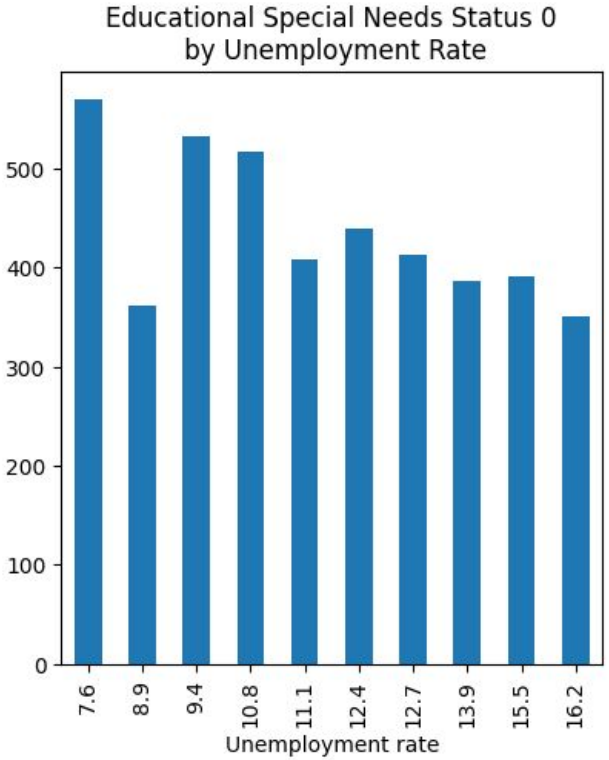
- From these statistics we infer:
  - Those with debtor status 1 show a higher proportion within the higher GDP category

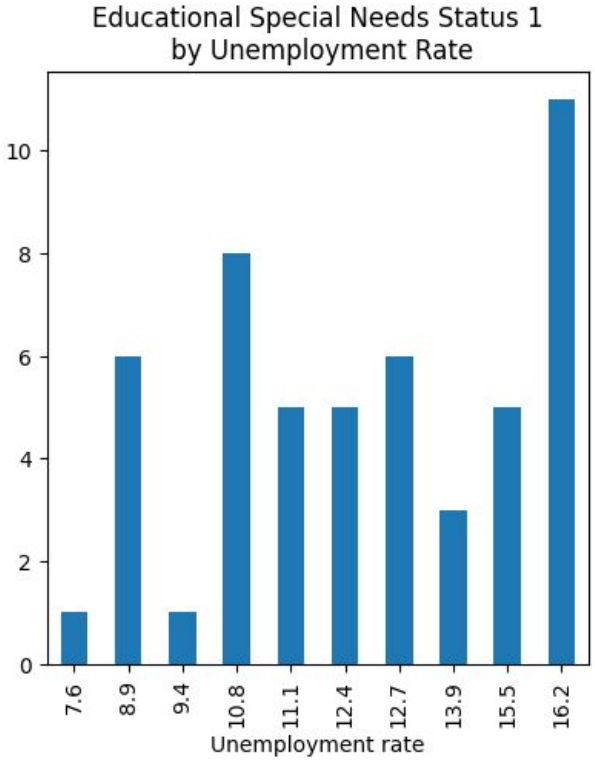




Educational Special Needs Status by Unemployment Rate

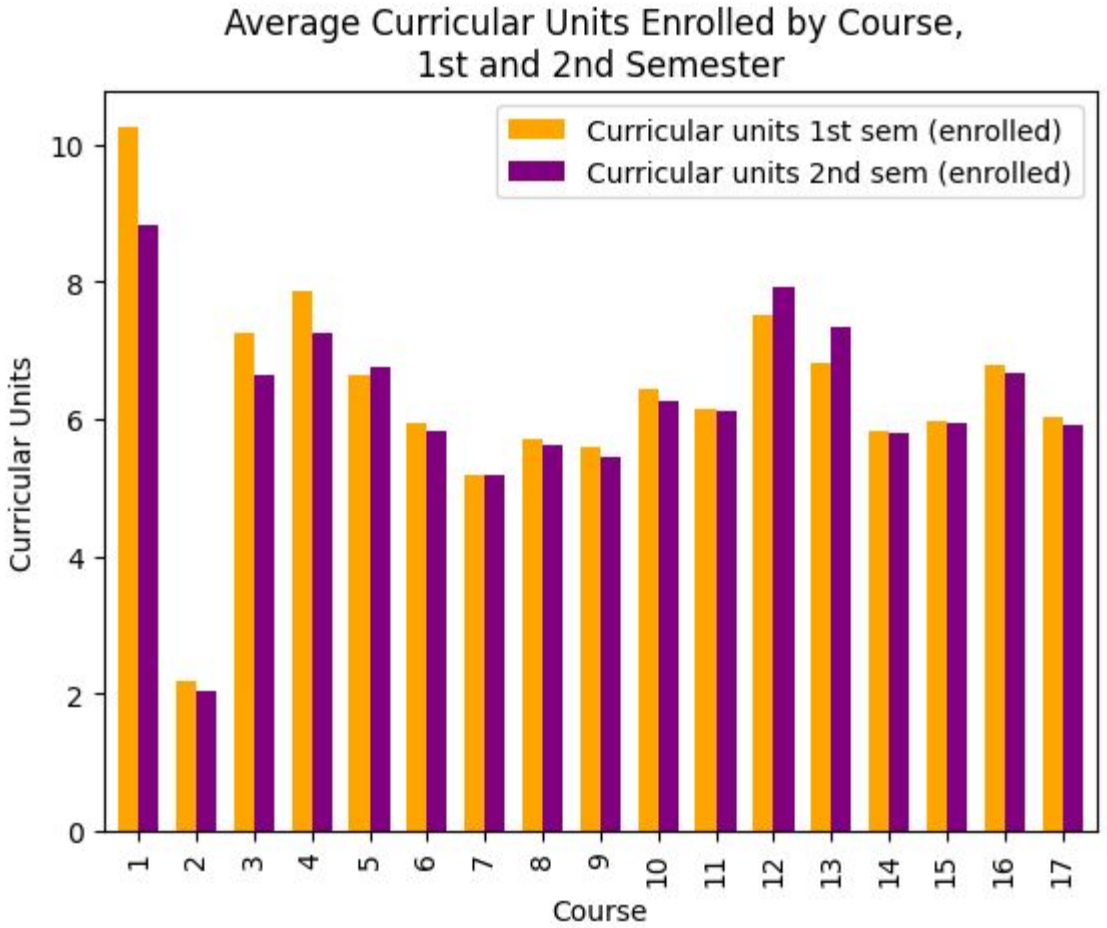
- From these statistics we infer:
  - Those with educational special needs status 0 have a lower and more evenly distributed unemployment rate
  - Those with educational special needs status 1 have a larger proportion of higher unemployment rates





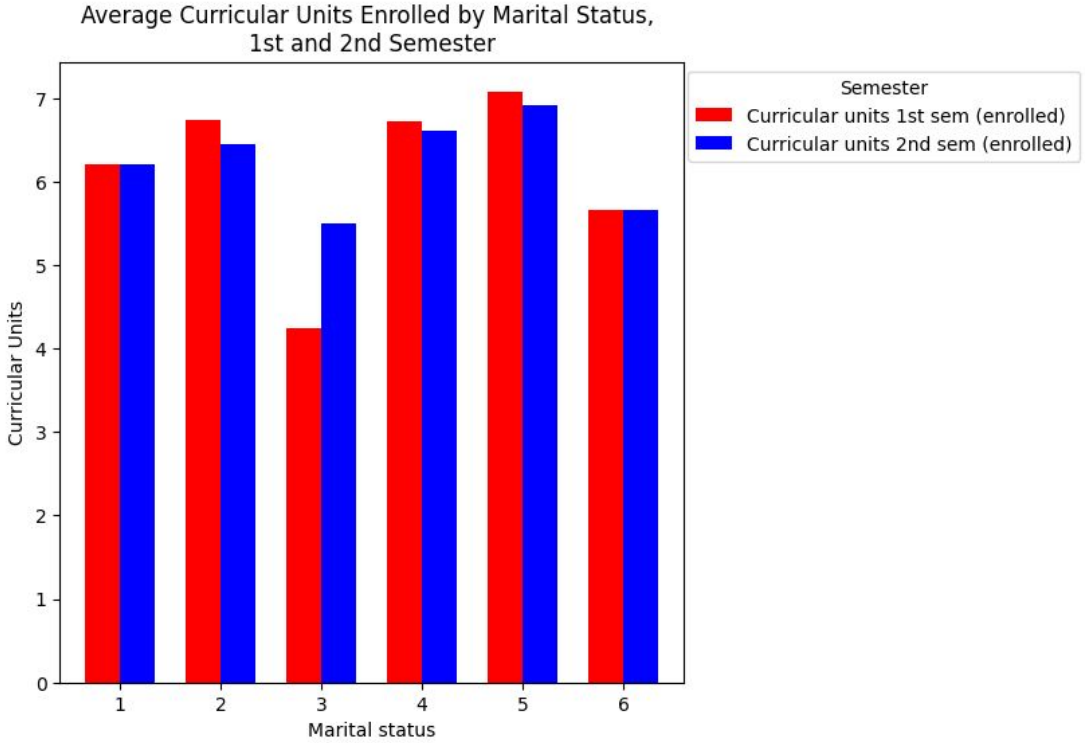
Curricular Units Enrolled (Average) by Course

- From these statistics we infer:
  - The higher average curricular units enrolled by course are in course 1
  - The lowest average curricular units enrolled by course are in course 2
  - Average units enrolled between first and second semester show no major differences



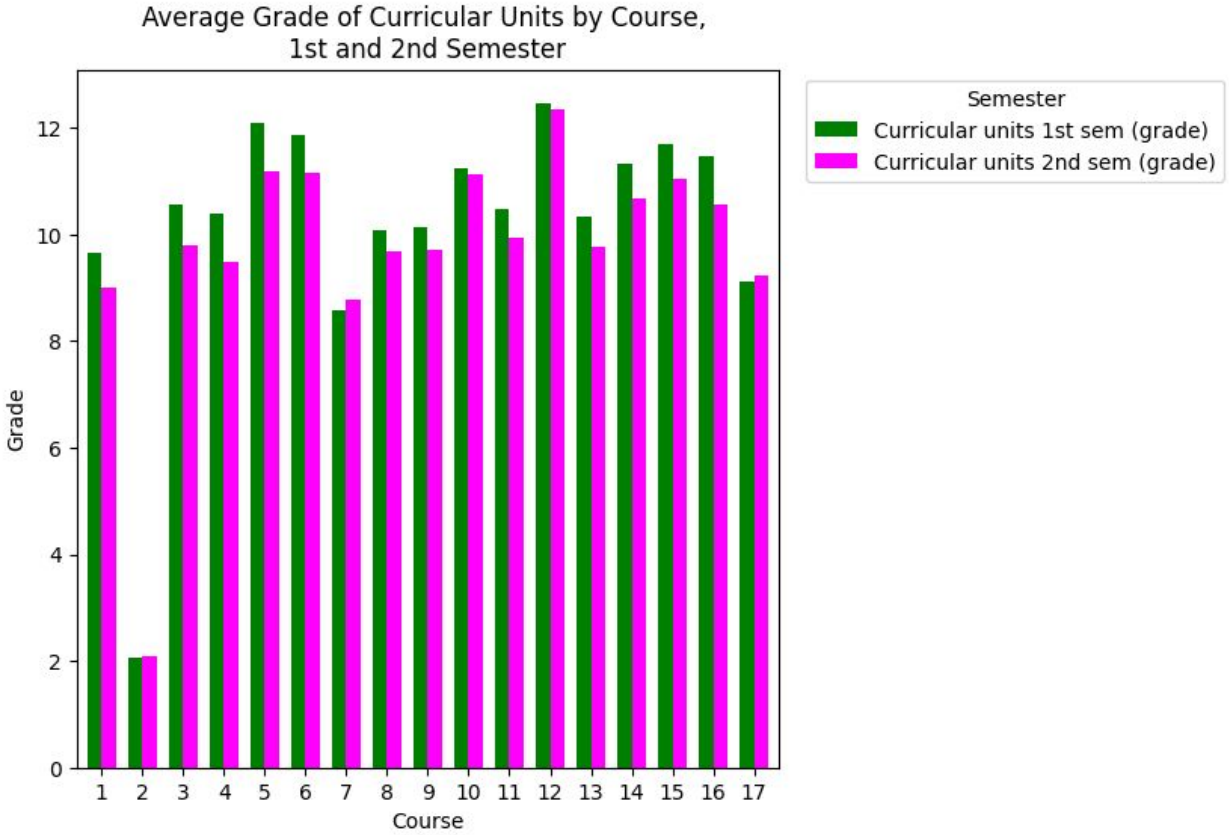
Curricular Units Enrolled (Average) by Marital Status

- From these statistics we infer:
  - There is a relatively even average of curricular units enrolled between semesters and marital statuses
  - However, marital status 3 shows a drop in the average first semester units enrolled, compared to average second semester enrollment



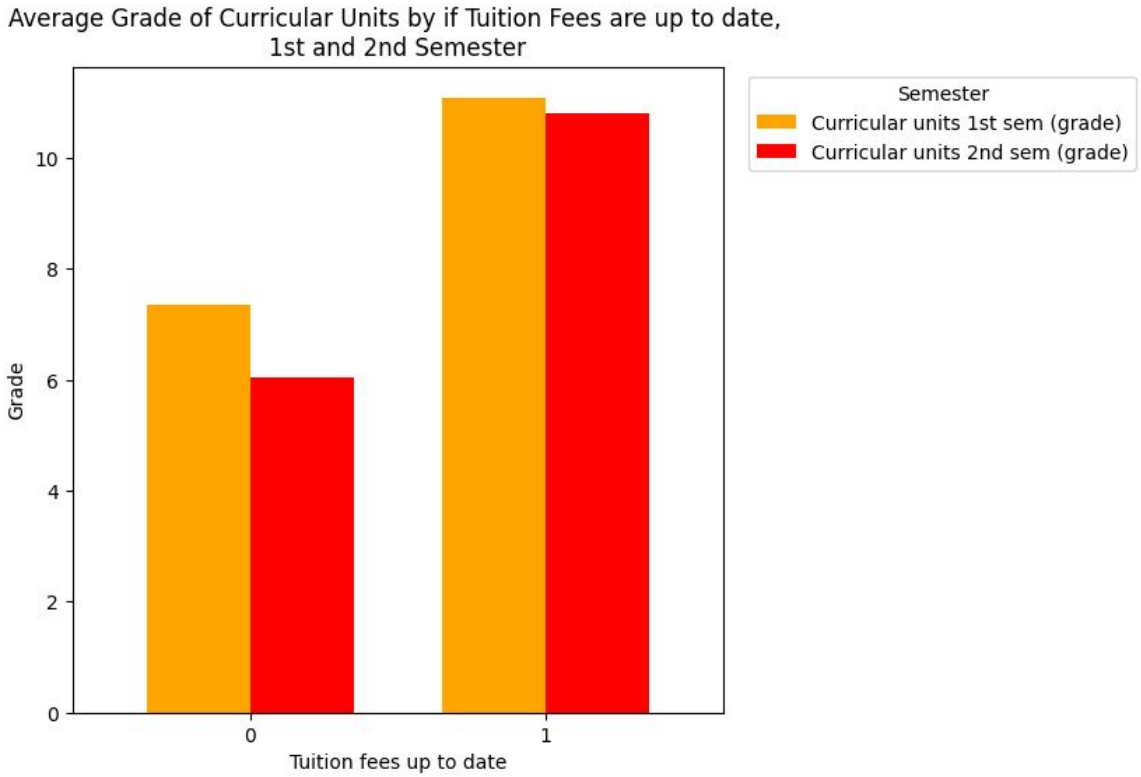
Curricular Units Grade (Average) by Course

- From these statistics we infer:
  - The grades received per course are mostly consistent between first and second semesters
  - The highest average grades are in course 12
  - The lowest average grades are in course 2



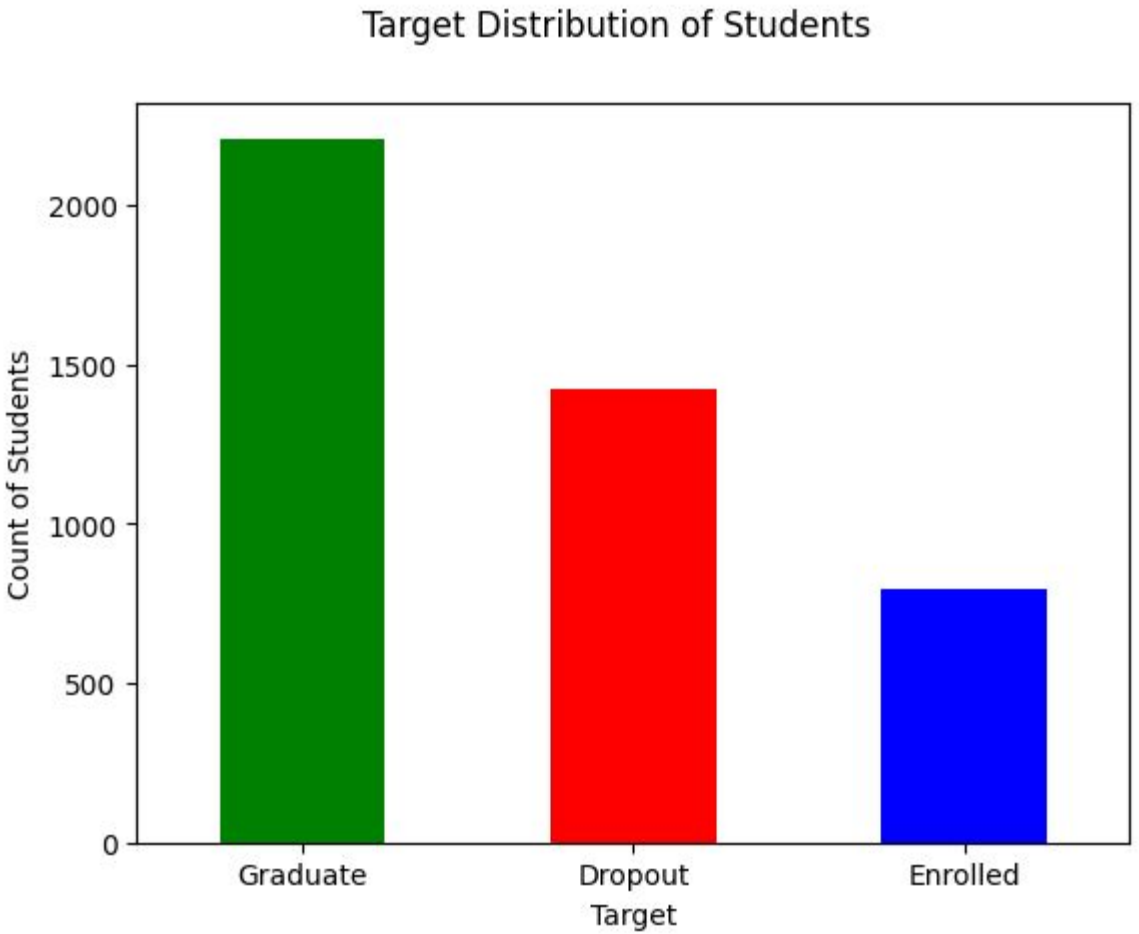
Curricular Units Grade (Average) by Tuition Fees Up to Date Status

- From these statistics we infer:
  - Those of tuition fees up to date status 0 received lower average grades between first and second semesters



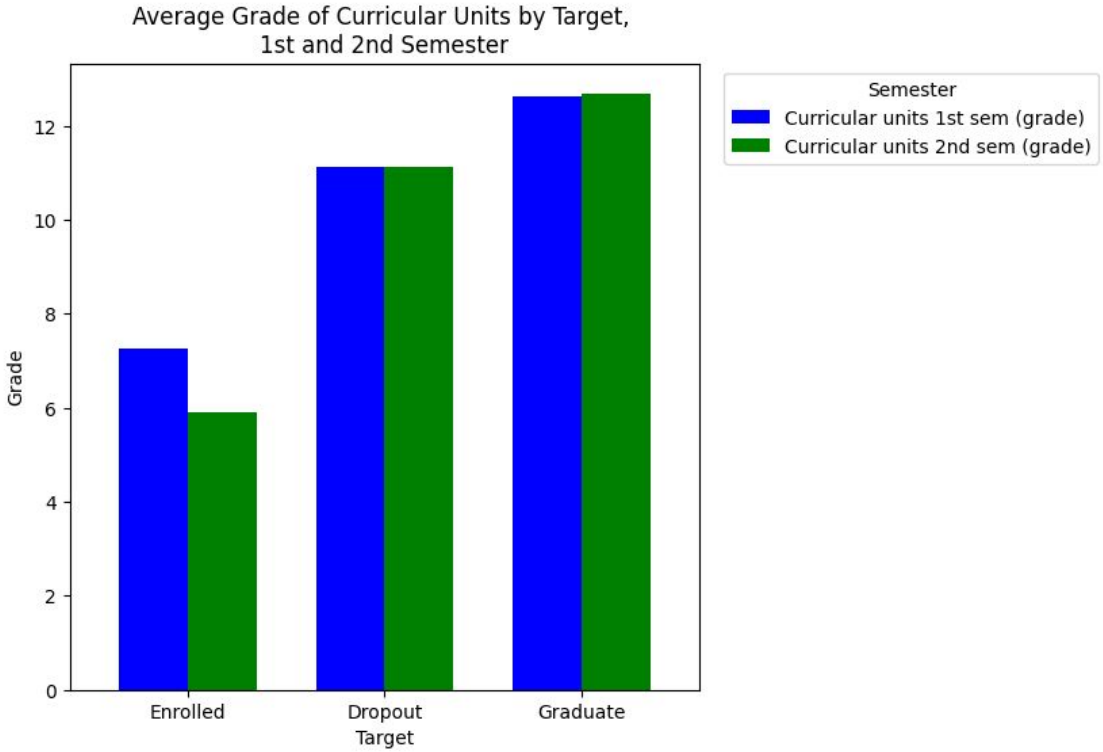
Target EDA

- Our target helps us further infer trends among the student population by observing their graduate, dropout, and enrolled status
- Here, we see that graduate occurs most frequently, dropout second, and enrolled third



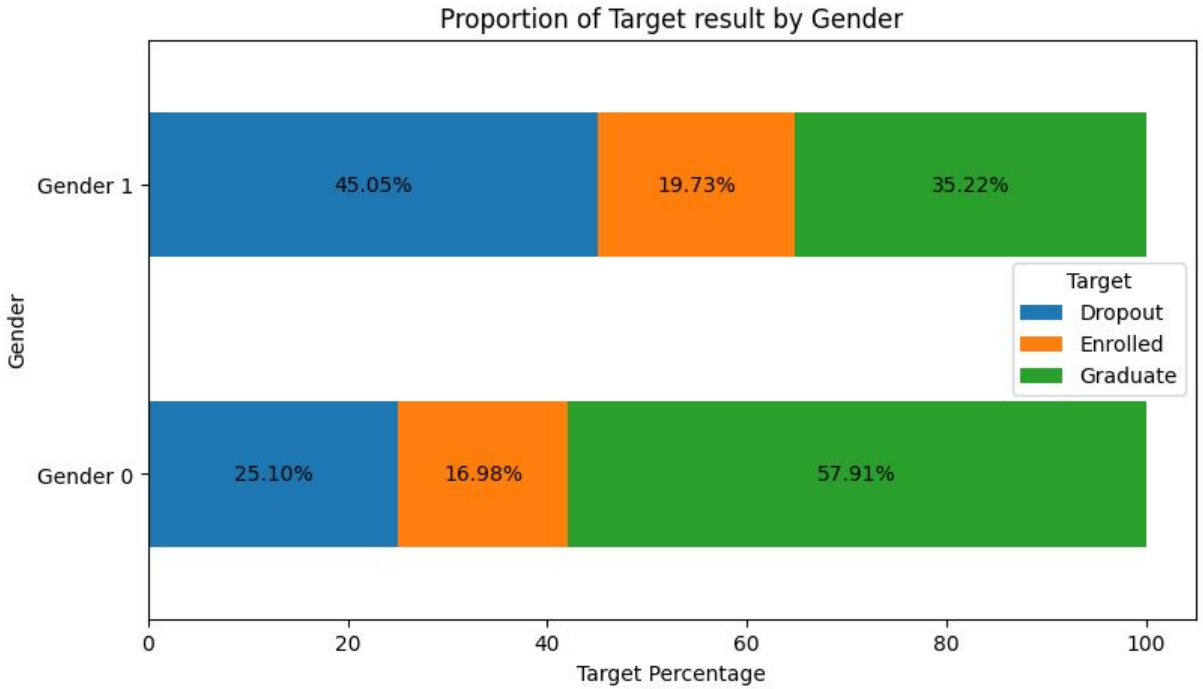
Target EDA (continued)

- From these statistics we infer:
  - Those with the target of dropout and graduate have higher average grades between semesters than those who are of the enrolled status



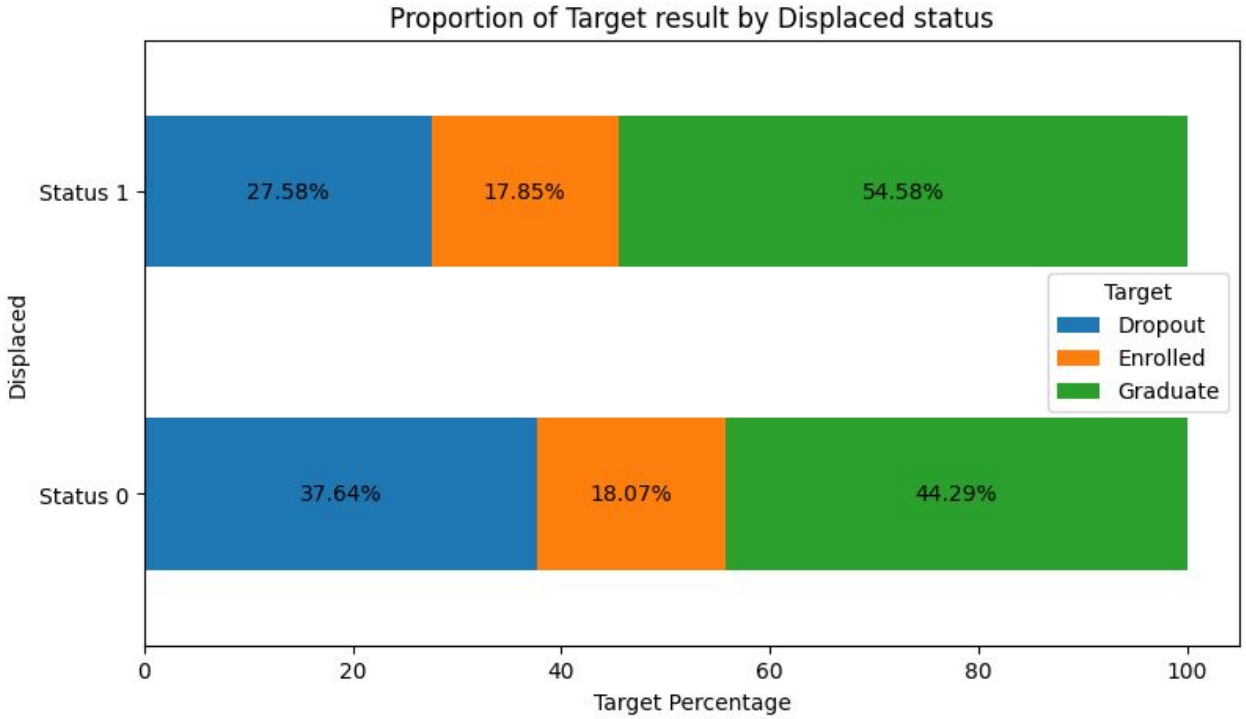
Target EDA (continued)

- We also infer:
  - A higher proportion of gender 1 represent the dropout target
  - A higher proportion of gender 0 represent the graduate target



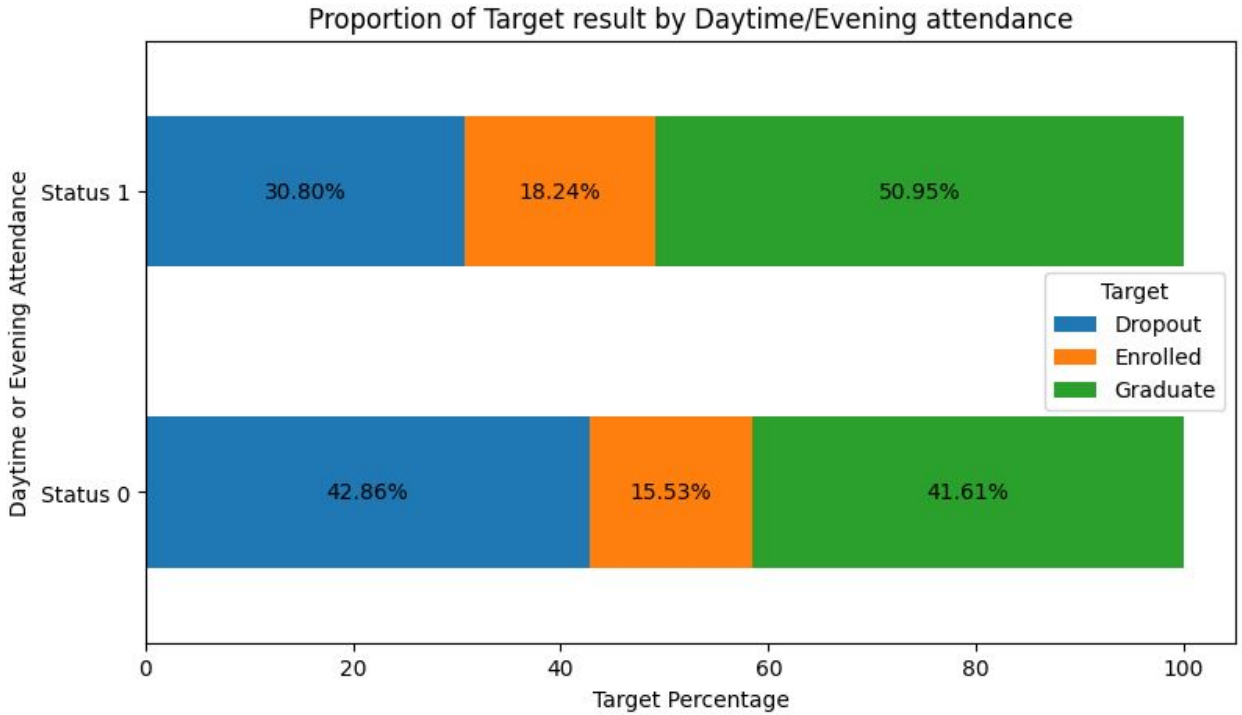
Target EDA (continued)

- We also infer:
  - Those of displaced status 0 tend to have a larger proportion of dropout as their target
  - Those of displaced status 1 tend to have a larger proportion of graduate as their target



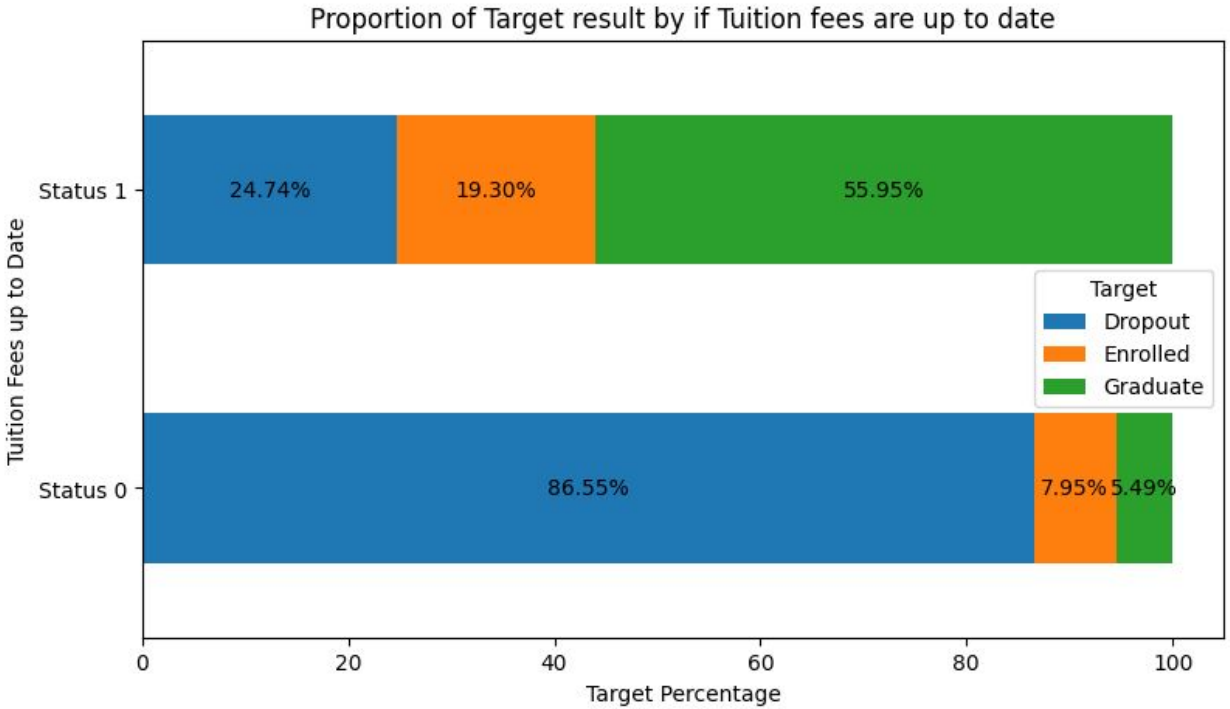
Target EDA (continued)

- We also infer:
  - A higher percentage of daytime/evening attendance status 1 are of the graduate target
  - A higher percentage of the daytime/evening attendance status 0 are of the dropout target



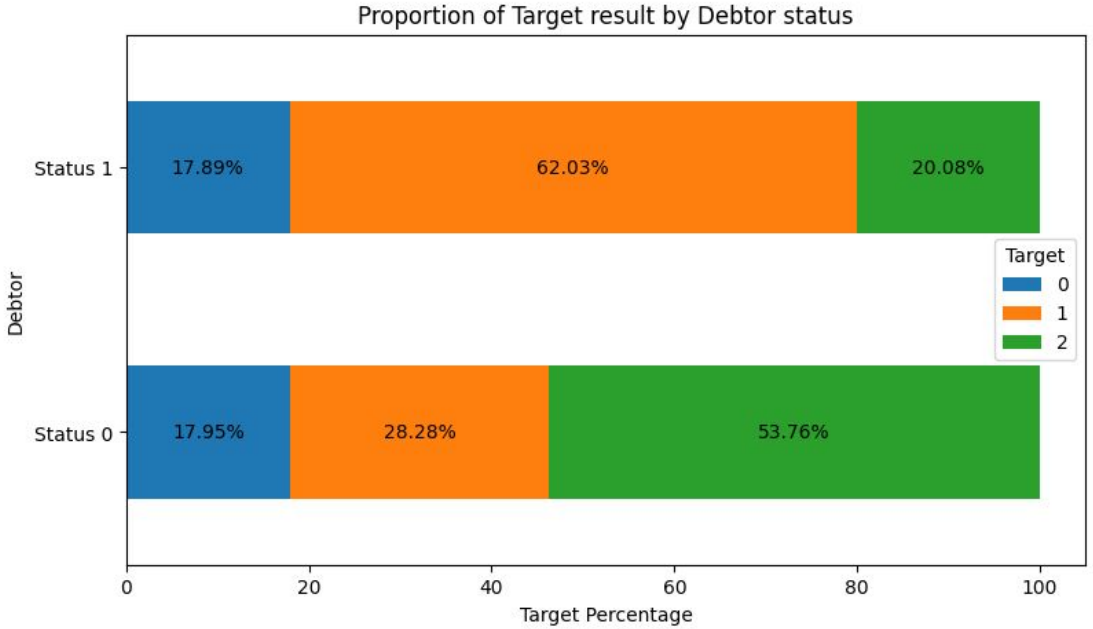
Target EDA (continued)

- We also infer:
  - A higher percentage of tuition fees up to date status 1 are of the graduate target
  - A very large majority of the tuition fees up to date status 0 are of the dropout category



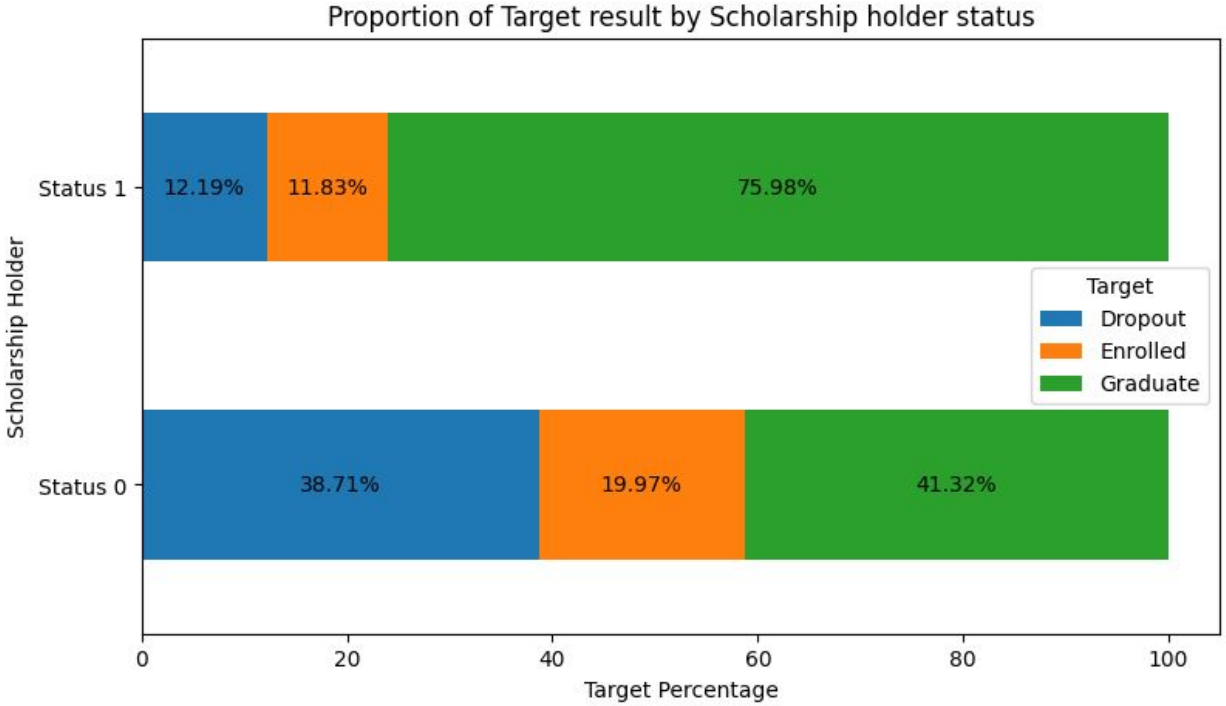
Target EDA (continued)

- We also infer:
  - The largest percentage of those with debtor status 1 are of the enrolled target
  - The largest percentage of debtor status 0 are of the graduate target



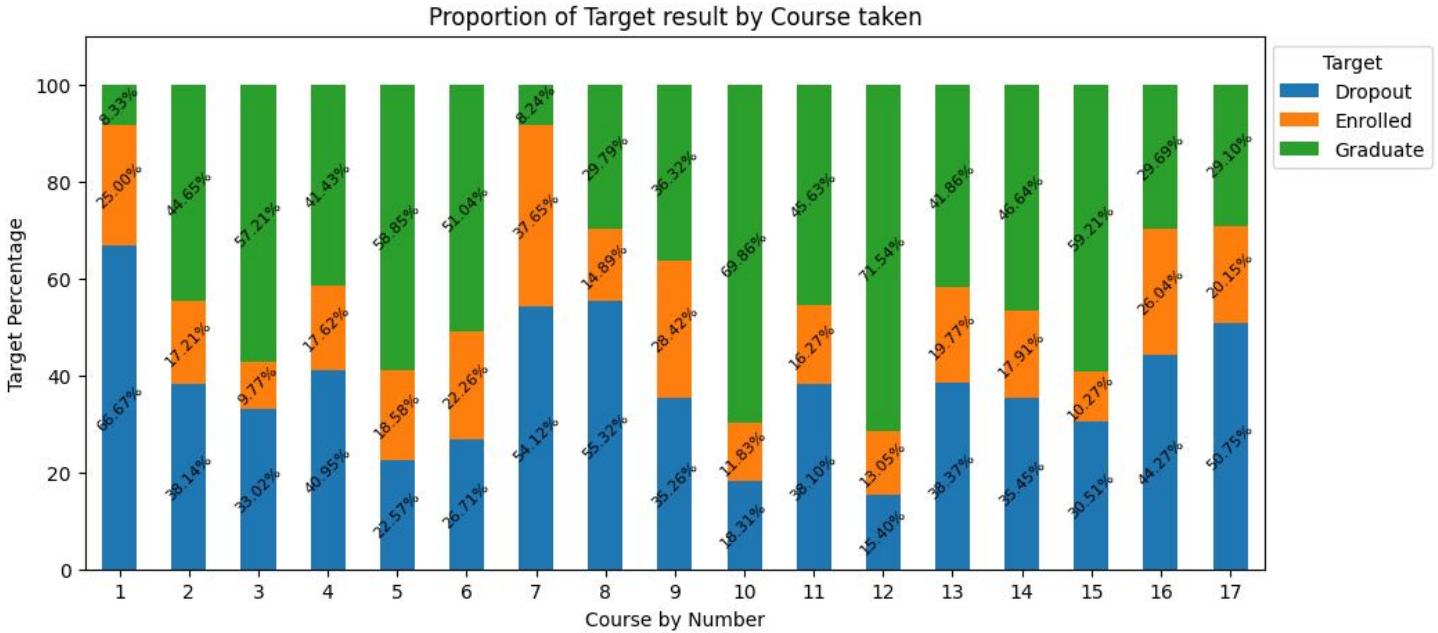
Target EDA (continued)

- We also infer:
  - A large majority of scholarship holder status 1 are of the graduate target
  - The percentage of graduate and dropout targets are split relatively evenly across scholarship holder status 0



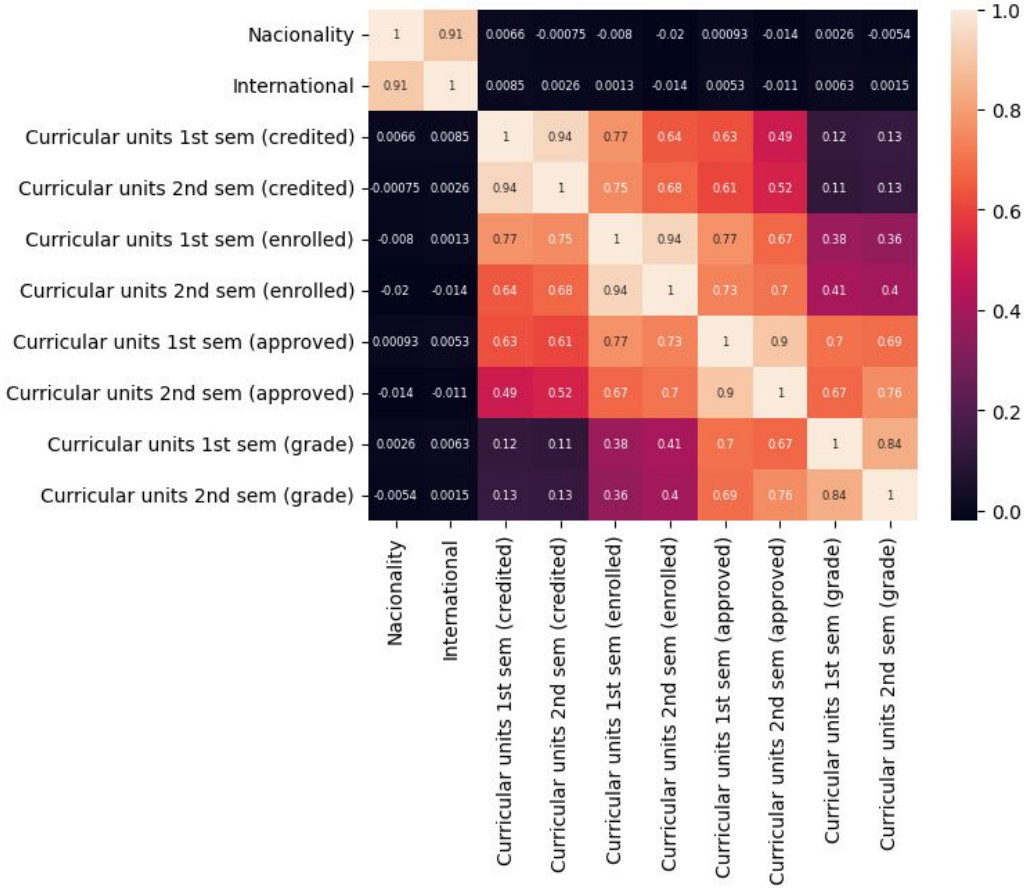
Target EDA (continued)

- We also infer:
  - The course with the highest rate of the dropout target is course number 1
  - The course with the highest rate of the graduate target is course number 10



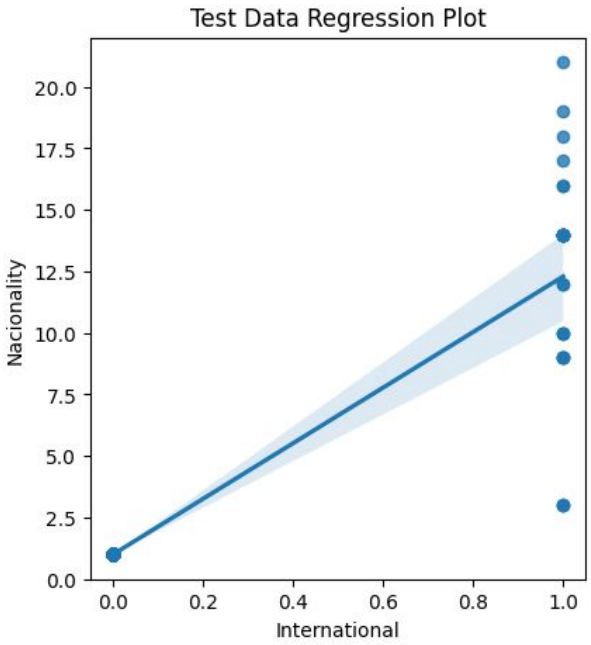
Correlation Between Variables and Predictive Modeling

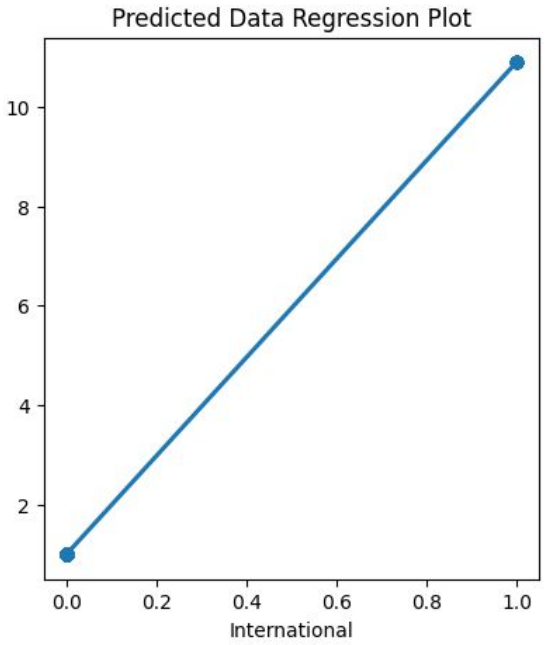
- Predictive modeling is used to show how the prediction of a future outcome can be made using input data
- Deciding what to model within this dataset was determined using the correlation coefficient
- The correlation coefficient displays the pairwise correlation between two variables of data compared, where the closer the number is to 1 or -1 showing a stronger linear relationship between the two variables
- The correlations shown on the right are those among the data that had a correlation coefficient higher than 0.80, which were then used within a Linear Regression model



Nationality/International Regression

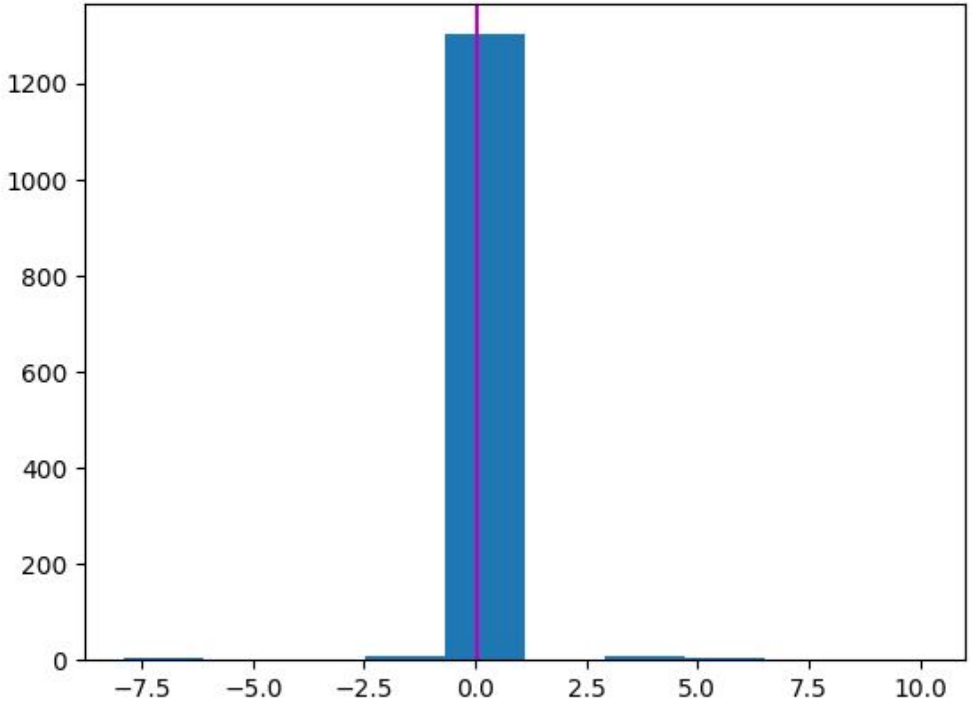
- This model shows the relationship between International Status and Nationality
- The linear regression on the left is that of the test data, and the linear regression on the right is that using the data that the model predicted
- The test data regression shows the predicted range of the values along the linear trendline, while also plotting the actual data points
- The predicted data regression shows the values that are those calculated, or approximated, using the model
- The r2 (R-Squared) value of the model shows how well the model explains the relationship between variables, how well the model creates approximated data results
    - A value closer to 1 means a greater fit, while a value closer to 0 means a poor fit
- The r2 value of this model is 0.85





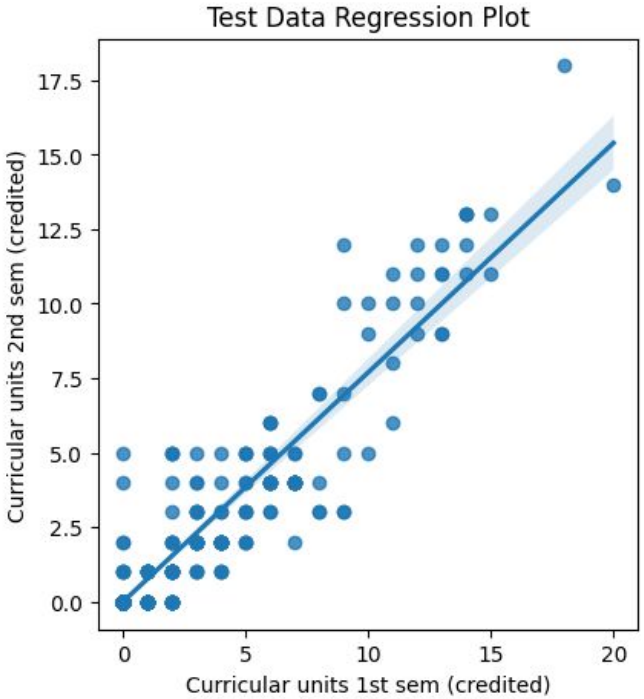
Nationality/International Regression (continued)

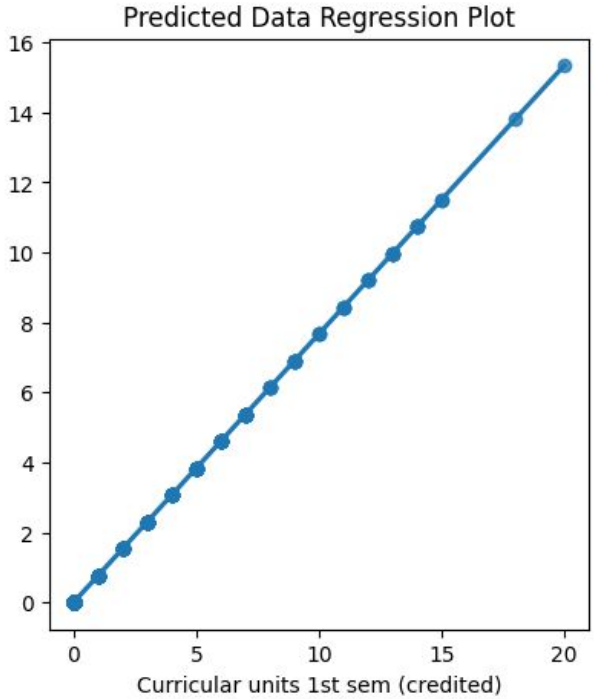
- The plot of the residuals also shows us how well the model fits the data
- They show the plotted differences between the test data outcome and the predicted data outcome
- The histogram shows that the residuals are normally distributed around 0, with the magenta line indicating the mean of the values
- A normal distribution shows that the model captures the main patterns and variation in the data, and that errors are independent and random



1st and 2nd Semester Units Credited Regression

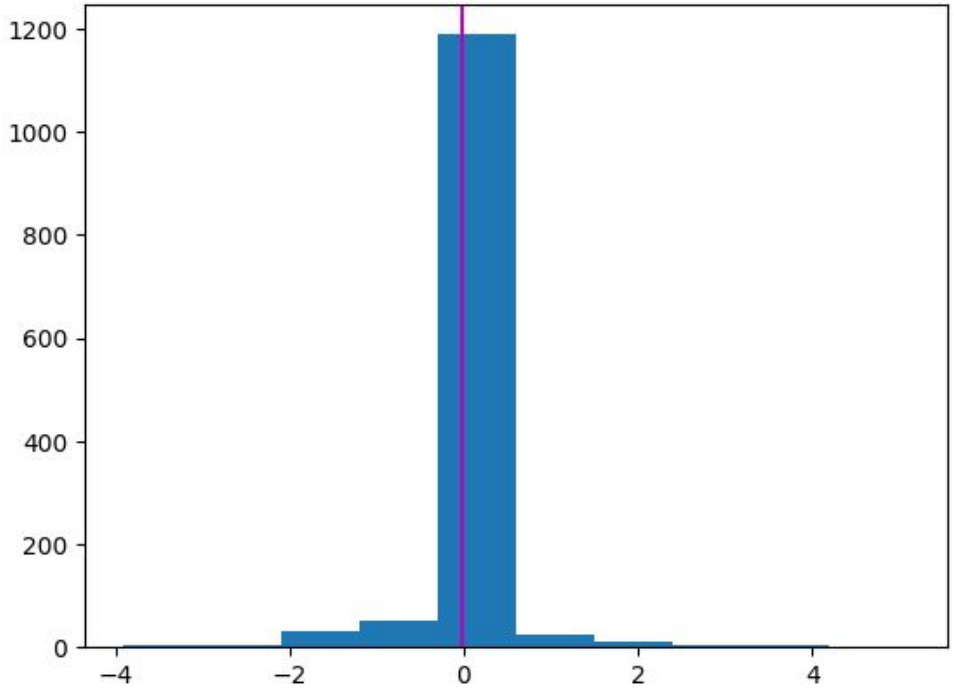
- The data and predicted data regressions here show the relationship between the 1st and 2nd semester curricular units credited
- The r2 value of the model is 0.90





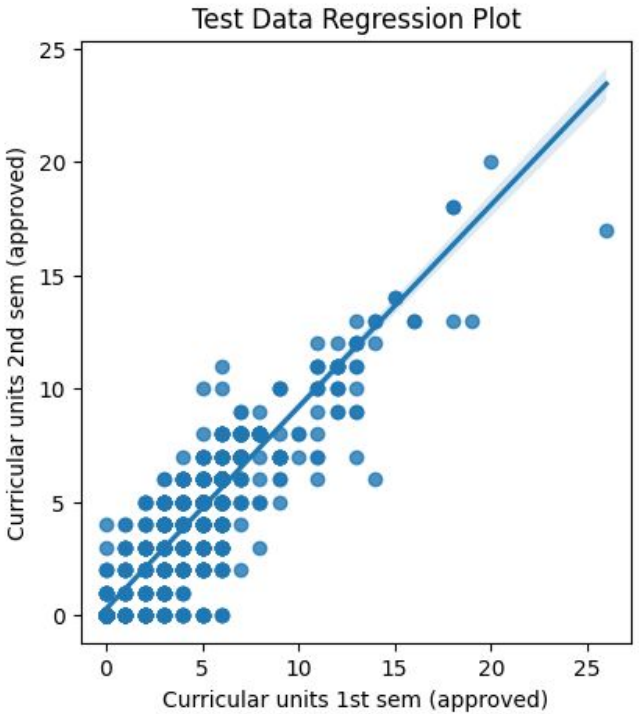
1st and 2nd Semester Units Credited Regression (continued)

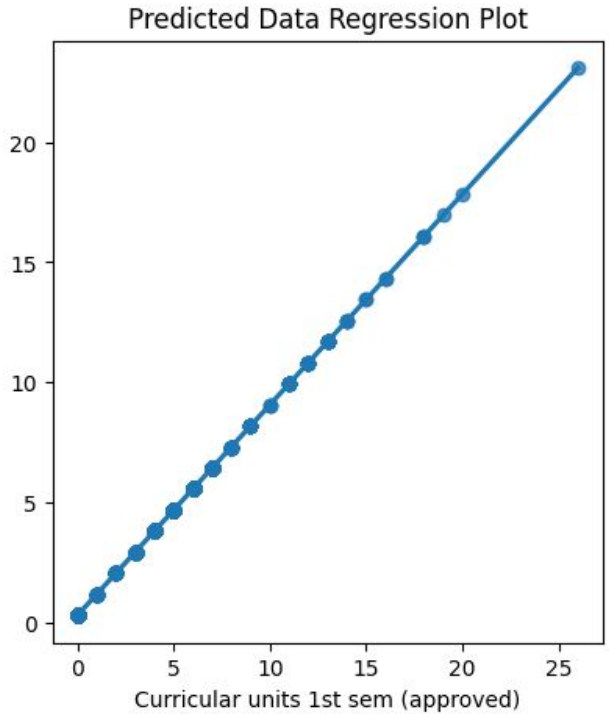
- The histogram shows that the residuals are normally distributed around 0, with the magenta line indicating the mean of the values



1st and 2nd Semester Units Approved Regression

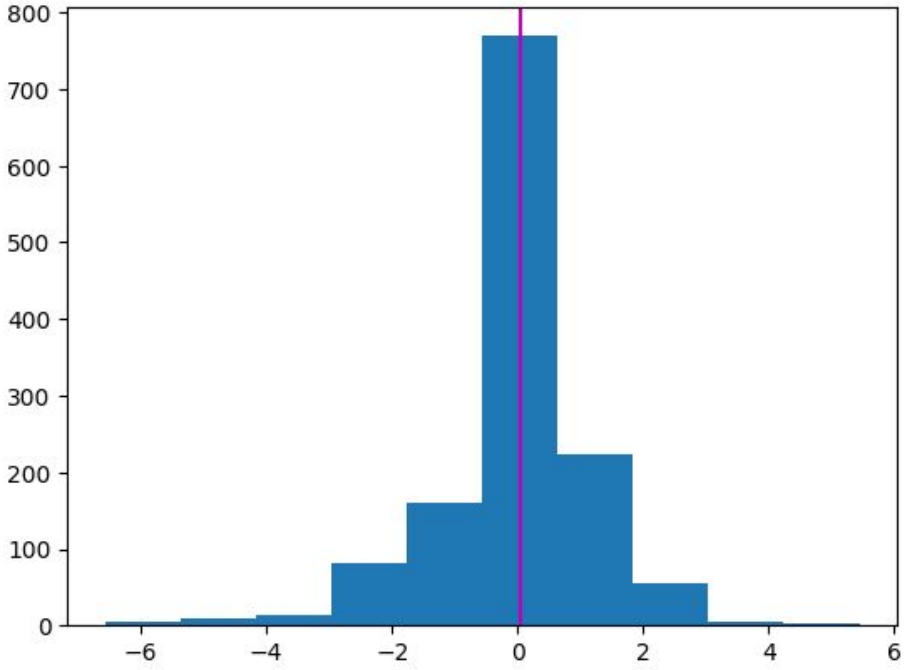
- The data and predicted data regressions here show the relationship between the 1st and 2nd semester curricular units approved
- The r2 value of the model is 0.82





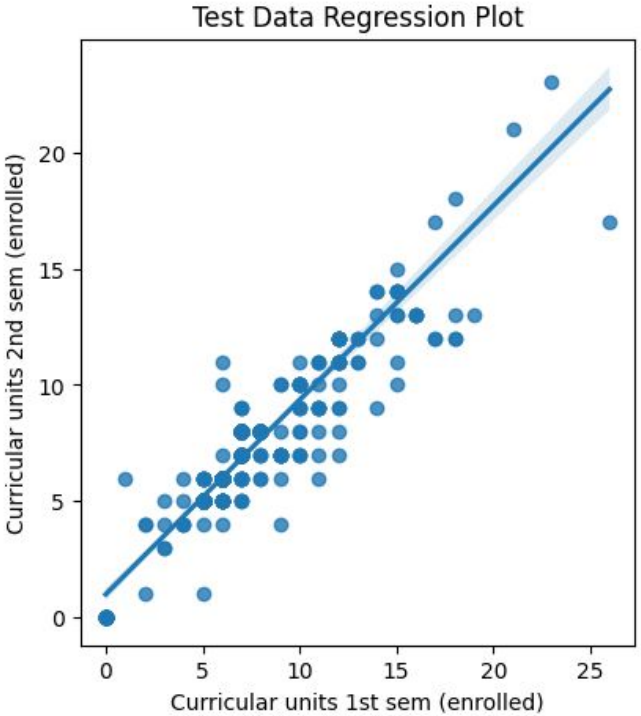
1st and 2nd Semester Units Approved Regression (continued)

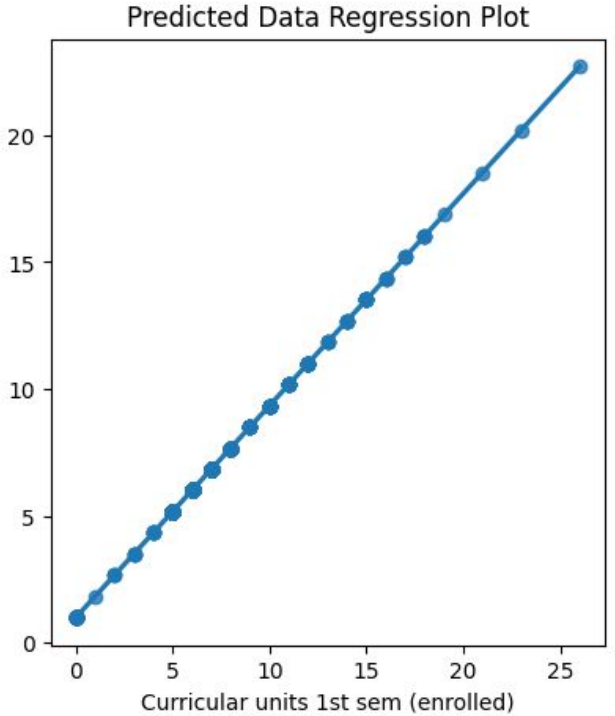
- The histogram shows that the residuals are normally distributed around 0, with the magenta line indicating the mean of the values



1st and 2nd Semester Units Enrolled Regression

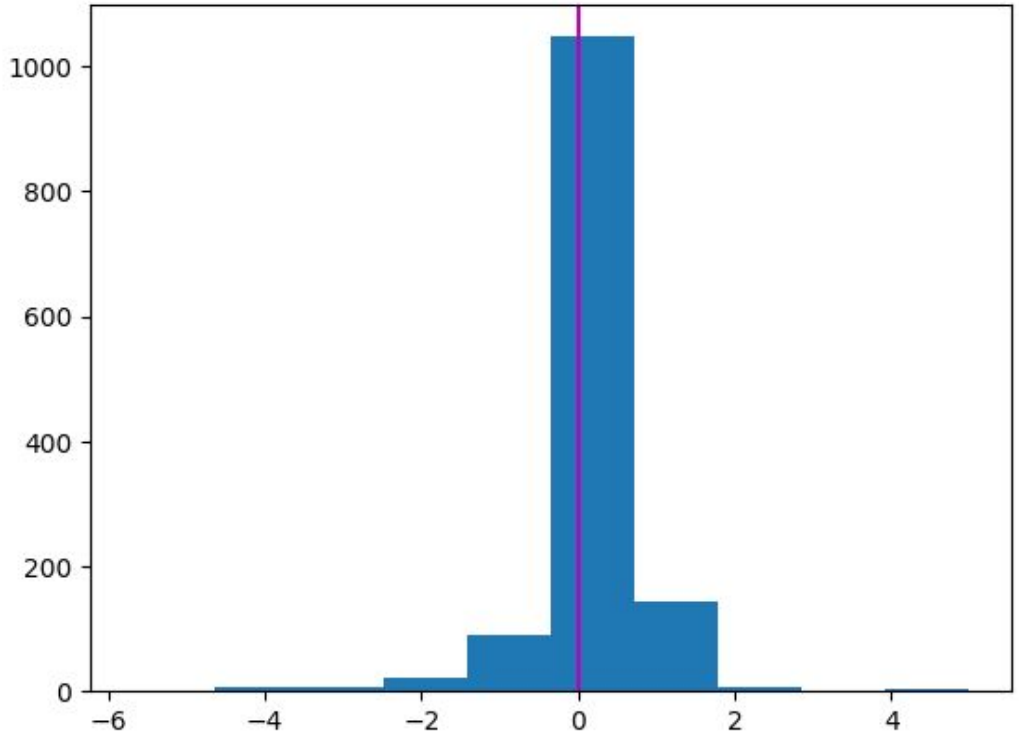
- The data and predicted data regressions here show the relationship between the 1st and 2nd semester curricular units enrolled
- The r2 value of the model is 0.88





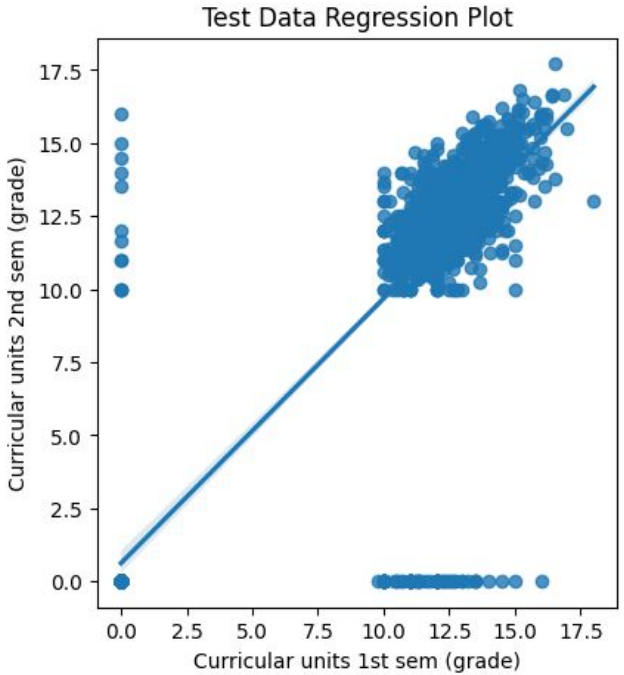
1st and 2nd Semester Units Enrolled Regression (continued)

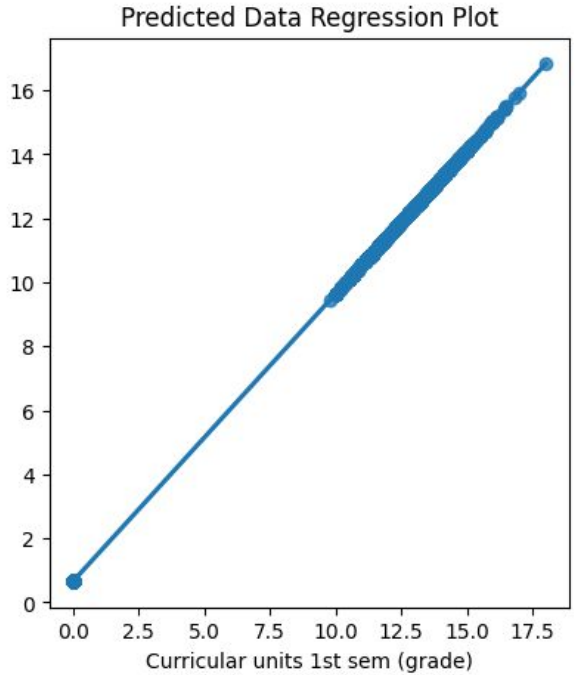
- The histogram shows that the residuals are normally distributed around 0, with the magenta line indicating the mean of the values



1st and 2nd Semester Units Grade Regression

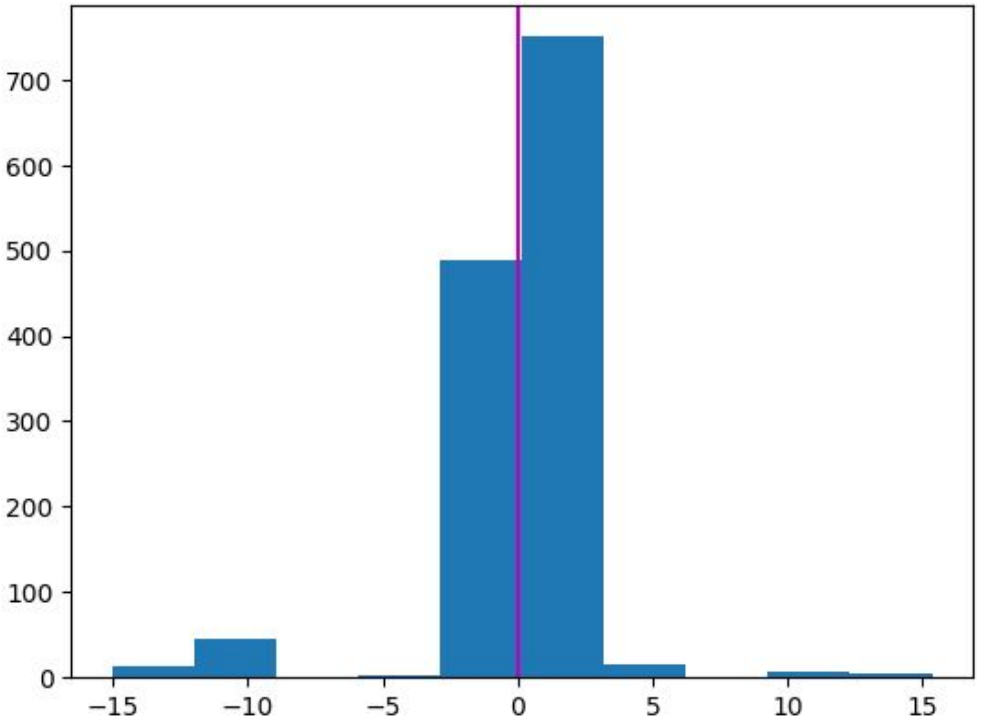
- The data and predicted data regressions here show the relationship between the grade of the 1st and 2nd semester curricular units
- The r2 value of the model is 0.70





1st and 2nd Semester Units Grade Regression (continued)

- The histogram shows that the residuals are normally distributed around 0, with the magenta line indicating the mean of the values



Link to GitHub