Import necessary libraries and # Load the dataset

```
# Import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Load the dataset with the correct delimiter
data_path = '../data/raw/data.csv'  # Adjust this path as necessary
df = pd.read_csv(data_path, delimiter=';')  # Specify semicolon as the delimiter
```

```
# Display the first few rows
df.head()
```

	Marital status	Application mode	Application order	Course	Daytime/evening attendance\t	Previous qualification	Previ qualificat (gra
0	1	17	5	171	1	1	12
1	1	15	1	9254	1	1	16
2	1	1	5	9070	1	1	12
3	1	17	2	9773	1	1	12
4	2	39	1	8014	0	1	10

5 rows × 37 columns

```
# Check the shape and data types
print(f"Dataset shape: {df.shape}")
print(df.dtypes)
```

```
Dataset shape: (4424, 37)
Marital status                                  int64
Application mode                                int64
Application order                               int64
Course                                          int64
Daytime/evening attendance\t                    int64
Previous qualification                          int64
Previous qualification (grade)                float64
Nacionality                                     int64
Mother's qualification                          int64
Father's qualification                          int64
Mother's occupation                             int64
Father's occupation                             int64
Admission grade                               float64
```

```
Displaced                                       int64
Educational special needs                       int64
Debtor                                          int64
Tuition fees up to date                         int64
Gender                                          int64
Scholarship holder                              int64
Age at enrollment                               int64
International                                   int64
Curricular units 1st sem (credited)             int64
Curricular units 1st sem (enrolled)             int64
Curricular units 1st sem (evaluations)          int64
Curricular units 1st sem (approved)             int64
Curricular units 1st sem (grade)              float64
Curricular units 1st sem (without evaluations)  int64
Curricular units 2nd sem (credited)             int64
Curricular units 2nd sem (enrolled)             int64
Curricular units 2nd sem (evaluations)          int64
Curricular units 2nd sem (approved)             int64
Curricular units 2nd sem (grade)              float64
Curricular units 2nd sem (without evaluations)  int64
Unemployment rate                             float64
Inflation rate                                float64
GDP                                           float64
Target                                         object
dtype: object
```

```
# Check the shape of the dataset
print(f"Dataset shape: {df.shape}")
```

```
Dataset shape: (4424, 37)
```

```
# Check the data types of each column
print(df.dtypes)
```

```
Marital status                                  int64
Application mode                                int64
Application order                               int64
Course                                          int64
Daytime/evening attendance\t                    int64
Previous qualification                          int64
Previous qualification (grade)                float64
Nacionality                                     int64
Mother's qualification                          int64
Father's qualification                          int64
Mother's occupation                             int64
Father's occupation                             int64
Admission grade                               float64
Displaced                                       int64
Educational special needs                       int64
Debtor                                          int64
Tuition fees up to date                         int64
Gender                                          int64
Scholarship holder                              int64
Age at enrollment                               int64
```

```
International                                     int64
Curricular units 1st sem (credited)             int64
Curricular units 1st sem (enrolled)             int64
Curricular units 1st sem (evaluations)          int64
Curricular units 1st sem (approved)             int64
Curricular units 1st sem (grade)              float64
Curricular units 1st sem (without evaluations)  int64
Curricular units 2nd sem (credited)             int64
Curricular units 2nd sem (enrolled)             int64
Curricular units 2nd sem (evaluations)          int64
Curricular units 2nd sem (approved)             int64
Curricular units 2nd sem (grade)              float64
Curricular units 2nd sem (without evaluations)  int64
Unemployment rate                             float64
Inflation rate                                float64
GDP                                           float64
Target                                         object
dtype: object
```

```
# Get summary statistics
df.describe()
```

	Marital status	Application mode	Application order	Course	Daytime/evening attendance\t	Previous qualificatior
count	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000
mean	1.178571	18.669078	1.727848	8856.642631	0.890823	4.577758
std	0.605747	17.484682	1.313793	2063.566416	0.311897	10.216592
min	1.000000	1.000000	0.000000	33.000000	0.000000	1.000000
25%	1.000000	1.000000	1.000000	9085.000000	1.000000	1.000000
50%	1.000000	17.000000	1.000000	9238.000000	1.000000	1.000000
75%	1.000000	39.000000	2.000000	9556.000000	1.000000	1.000000
max	6.000000	57.000000	9.000000	9991.000000	1.000000	43.000000

8 rows × 36 columns

```
df.describe(include=['object', 'category'])
```

	Target
count	4424
unique	3
top	Graduate
freq	2209

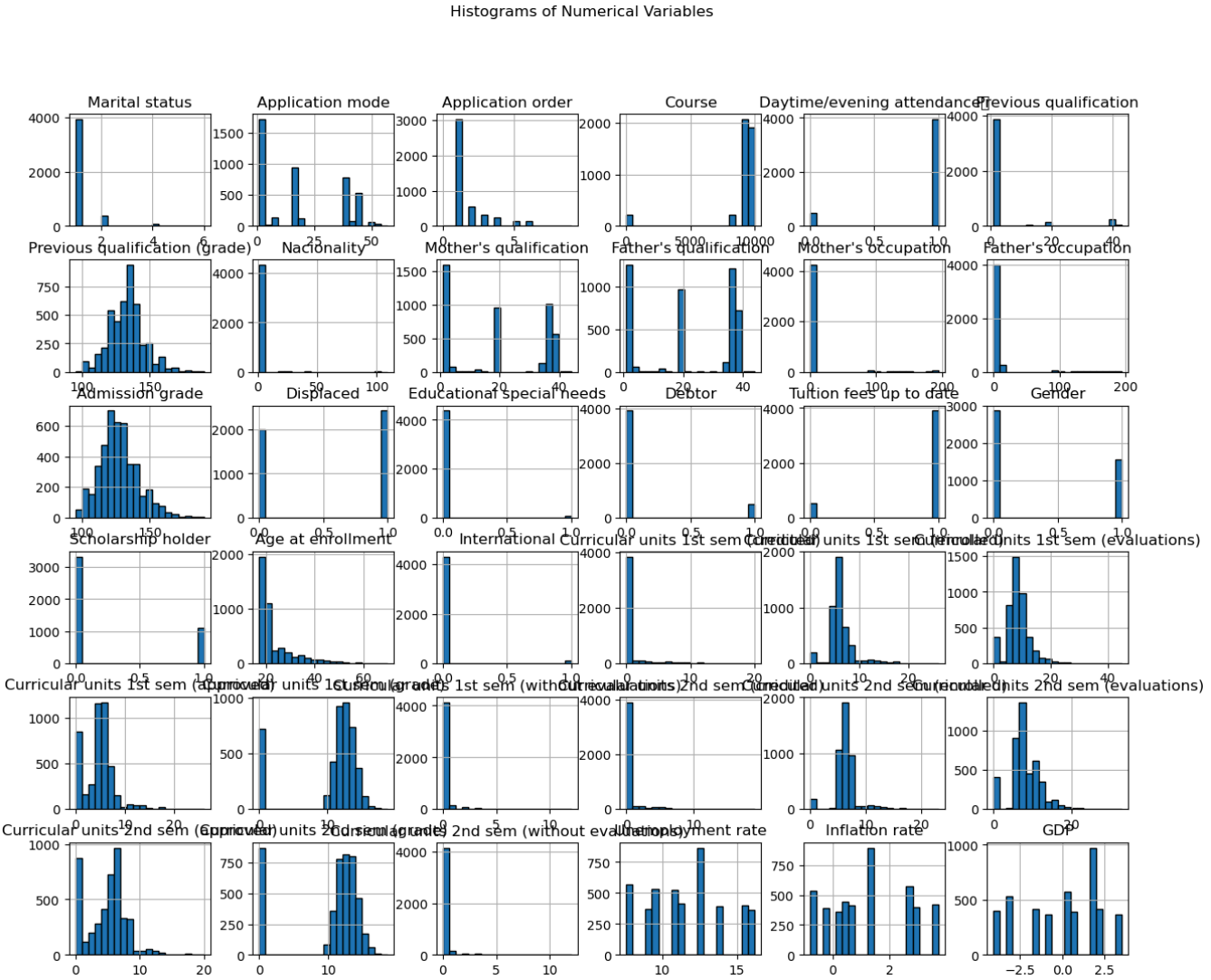
EDA

1.Univariate Analysis

Create histograms and box plots for numerical variables and bar charts for categorical variables, along with descriptive statistics.

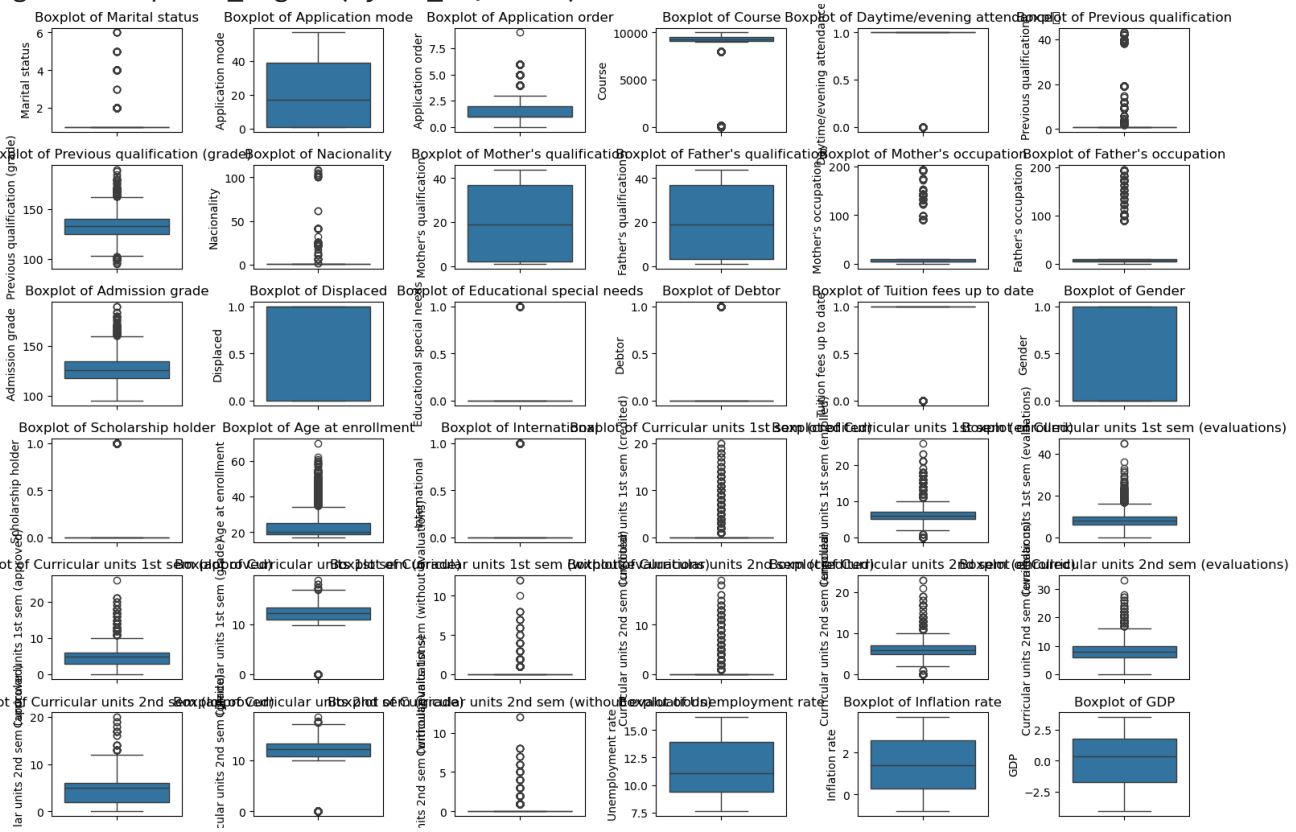
```
# Univariate Analysis for Numerical Variables (Histograms and Box Plots)
numerical_columns = df.select_dtypes(include=['int64', 'float64']).columns


# Histograms for numerical variables
df[numerical_columns].hist(figsize=(15, 12), bins=20, edgecolor='black')
plt.suptitle('Histograms of Numerical Variables')
plt.show()
```

```
c:\Users\matiwos.desalegn\AppData\Local\miniconda3\envs\student-dropout-prediction\lib\s
  fig.canvas.print_figure(bytes_io, **kw)
```

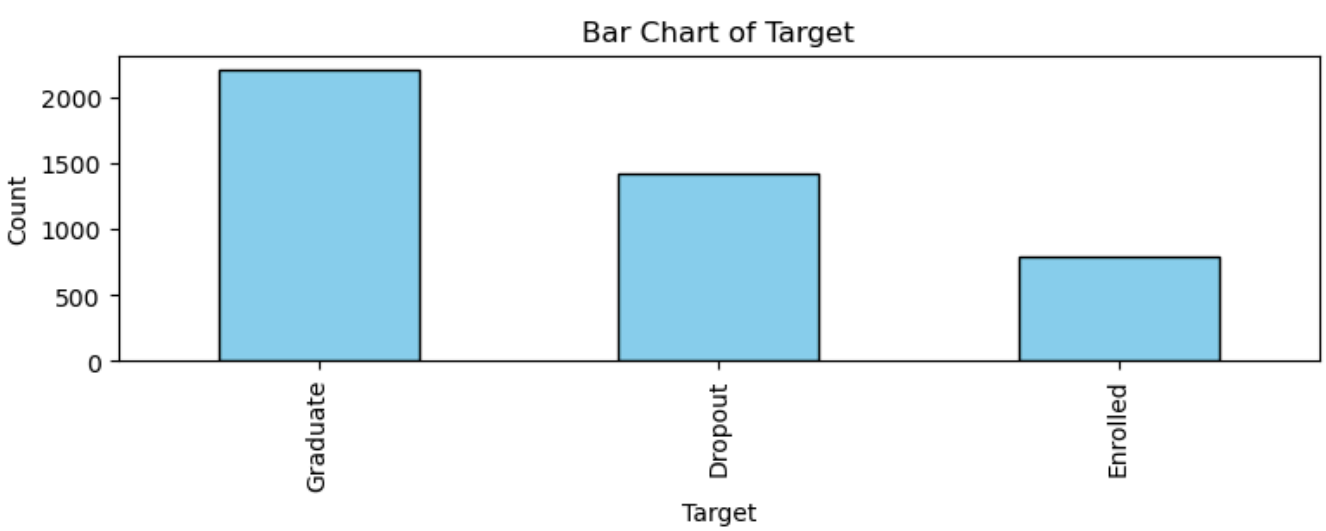



```
# Box plots for numerical variables
plt.figure(figsize=(15, 12))
for i, col in enumerate(numerical_columns):
    plt.subplot(7, 6, i+1)
    sns.boxplot(y=df[col])
    plt.title(f'Boxplot of {col}')
plt.tight_layout()
plt.show()
```

```
C:\Users\matiwos.desalegn\AppData\Local\Temp\ipykernel_45416\1009060163.py:7: UserWarnir
  plt.tight_layout()
c:\Users\matiwos.desalegn\AppData\Local\miniconda3\envs\student-dropout-prediction\lib\s
  fig.canvas.print_figure(bytes_io, **kw)
```




```
# Bar charts for categorical variables
categorical_columns = df.select_dtypes(include=['object']).columns

plt.figure(figsize=(15, 8))
for i, col in enumerate(categorical_columns):
    plt.subplot(3, 2, i+1)
    df[col].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title(f'Bar Chart of {col}')
    plt.ylabel('Count')
plt.tight_layout()
plt.show()
```




```
# Descriptive statistics for numerical variables
print(df[numerical_columns].describe())
```

```
50%                                    5.000000                       12.200000
75%                                    6.000000                       13.333333
max                                   20.000000                       18.571429

       Curricular units 2nd sem (without evaluations)  Unemployment rate  \
count                                     4424.000000        4424.000000
mean                                         0.150316          11.566139
std                                          0.753774           2.663850
min                                          0.000000           7.600000
25%                                          0.000000           9.400000
50%                                          0.000000          11.100000
75%                                          0.000000          13.900000
max                                         12.000000          16.200000

       Inflation rate          GDP
count     4424.000000  4424.000000
mean         1.228029     0.001969
std          1.382711     2.269935
min         -0.800000    -4.060000
25%          0.300000    -1.700000
50%          1.400000     0.320000
75%          2.600000     1.790000
max          3.700000     3.510000

[8 rows x 36 columns]
```

```
# Descriptive statistics for categorical variables
print(df[categorical_columns].describe())
```

```
          Target
count       4424
unique         3
top     Graduate
freq        2209
```

2. Bivariate Analysis

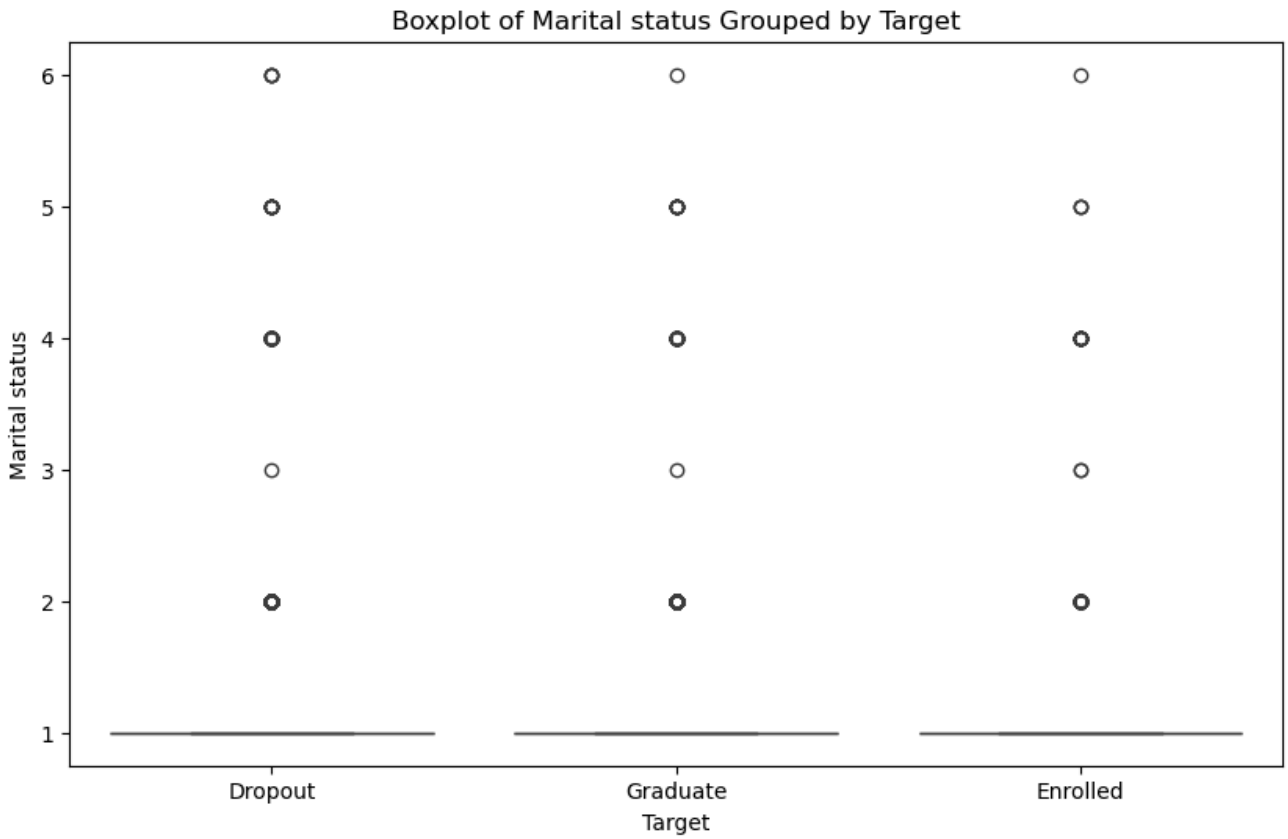
Create scatter plots, box plots, perform correlation analysis, and chi-square tests.

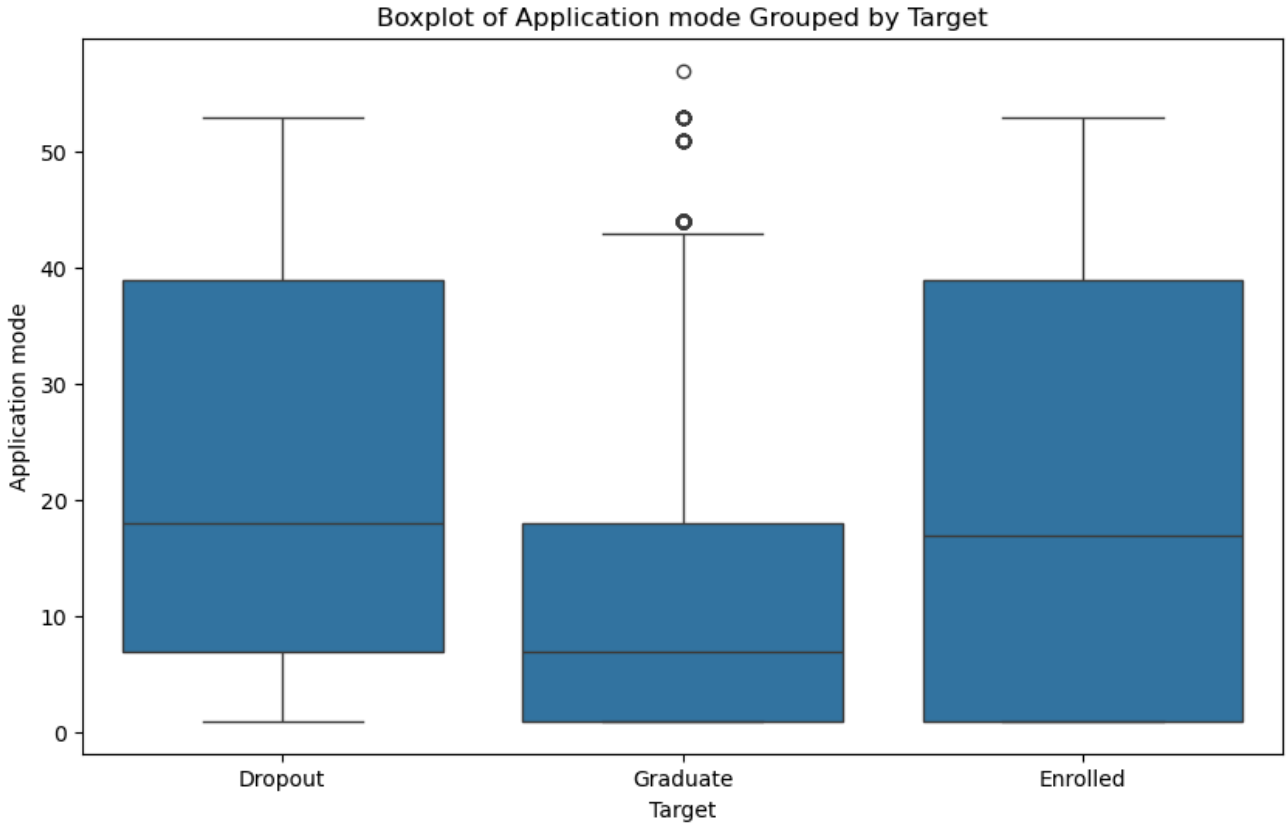
```
# Scatter plots for pairs of numerical variables
sns.pairplot(df[numerical_columns])
plt.show()
```

```
c:\Users\matiwos.desalegn\AppData\Local\miniconda3\envs\student-dropout-prediction\lib\s
  self._figure.tight_layout(*args, **kwargs)
c:\Users\matiwos.desalegn\AppData\Local\miniconda3\envs\student-dropout-prediction\lib\s
  fig.canvas.print_figure(bytes_io, **kw)
```

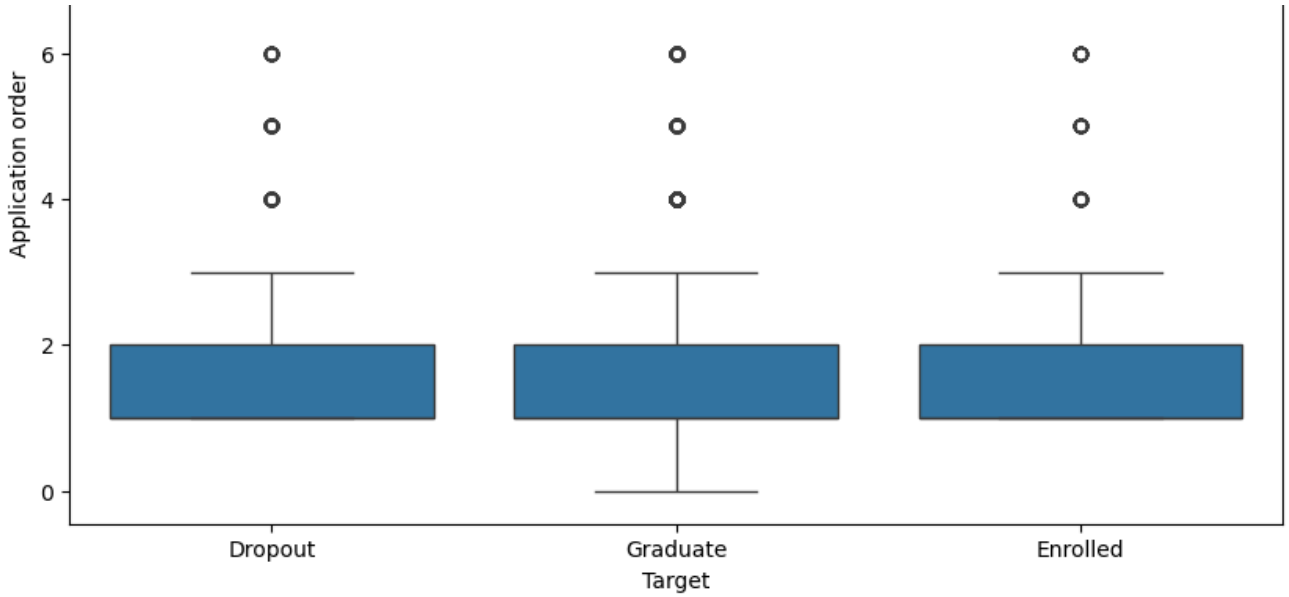
```
# Box plots of numerical variables grouped by categorical variables
for col in numerical_columns:
```

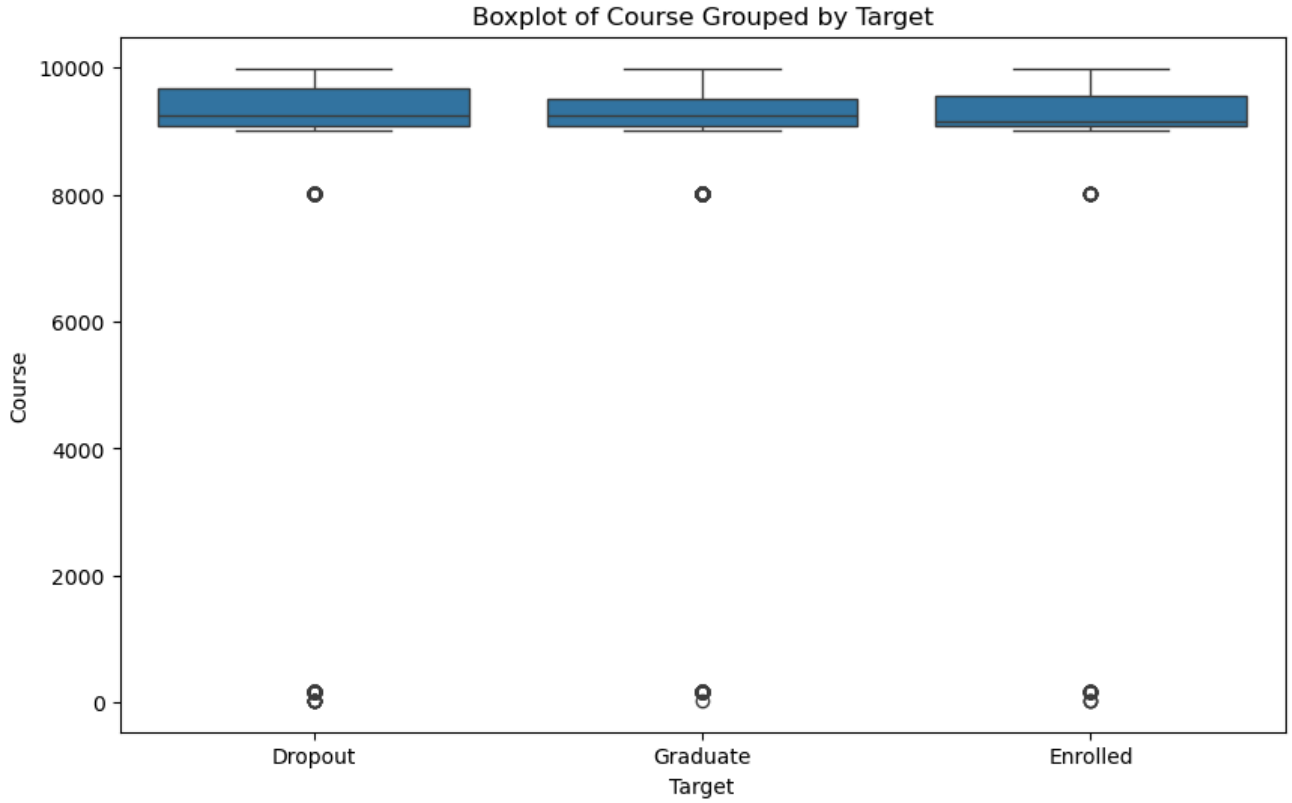
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Target', y=df[col], data=df)
plt.title(f'Boxplot of {col} Grouped by Target')
plt.show()
```


Boxplot of Marital status Grouped by Target
Marital status
1
2
3
4
5
6
Dropout
Graduate
Enrolled
Target

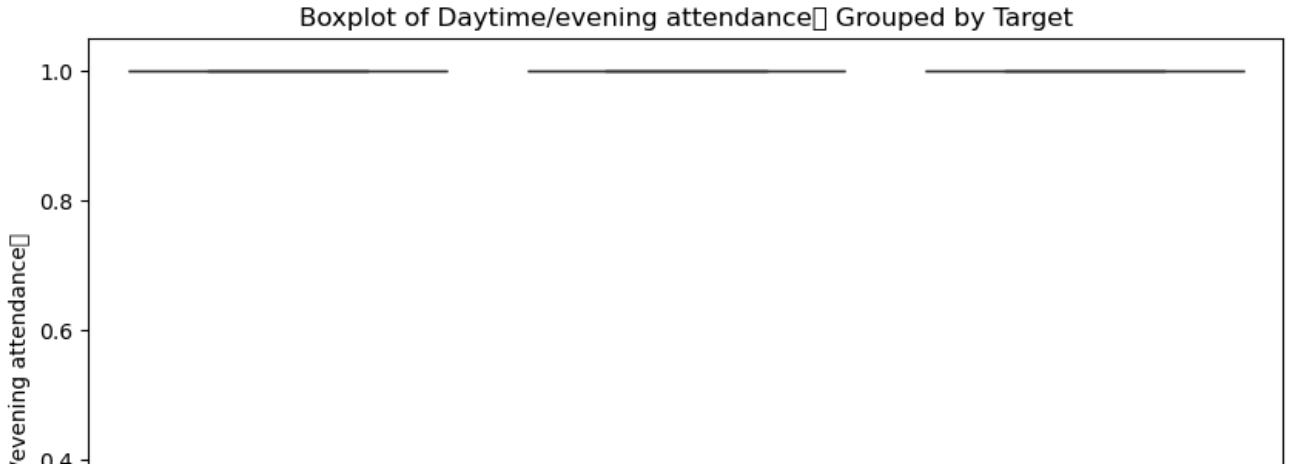

Boxplot of Application mode Grouped by Target
Application mode
0
10
20
30
40
50
Dropout
Graduate
Enrolled
Target

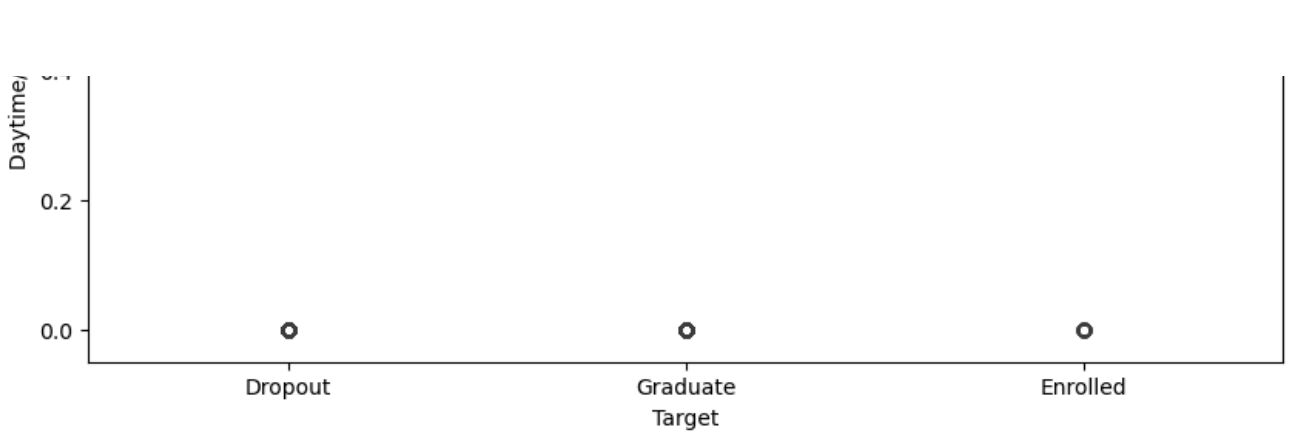

Boxplot of Application order Grouped by Target
8

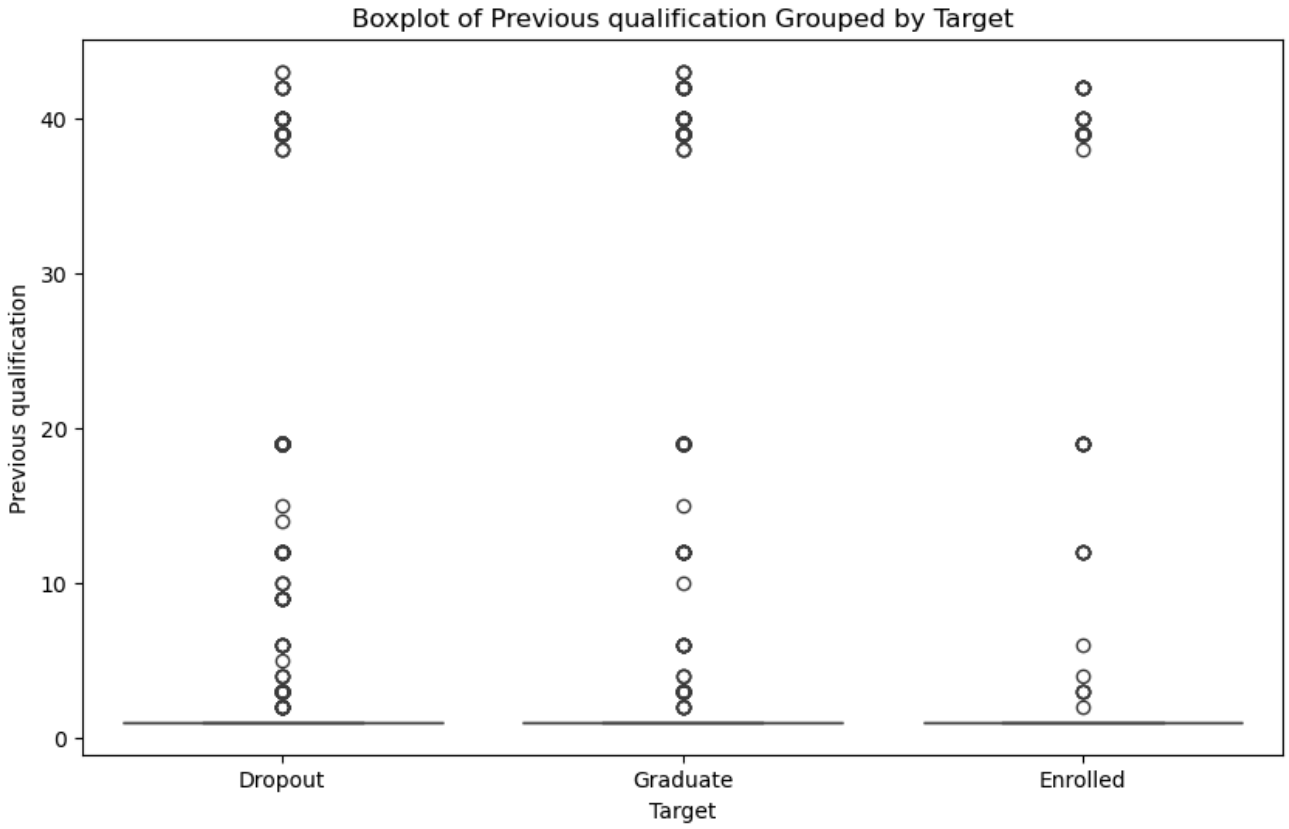


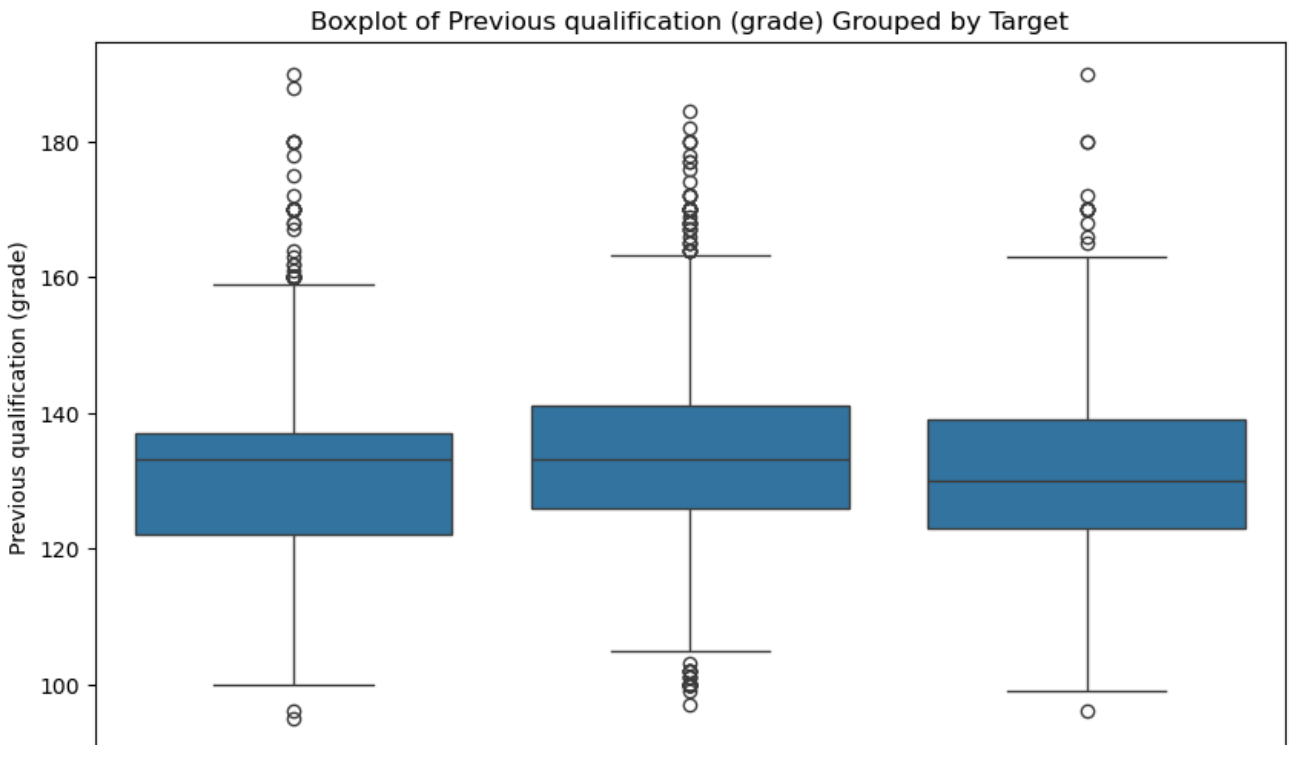


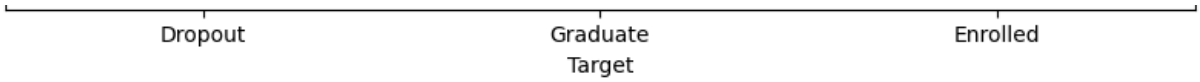

```
c:\Users\matiwos.desalegn\AppData\Local\miniconda3\envs\student-dropout-prediction\li
  fig.canvas.print_figure(bytes_io, **kw)
```

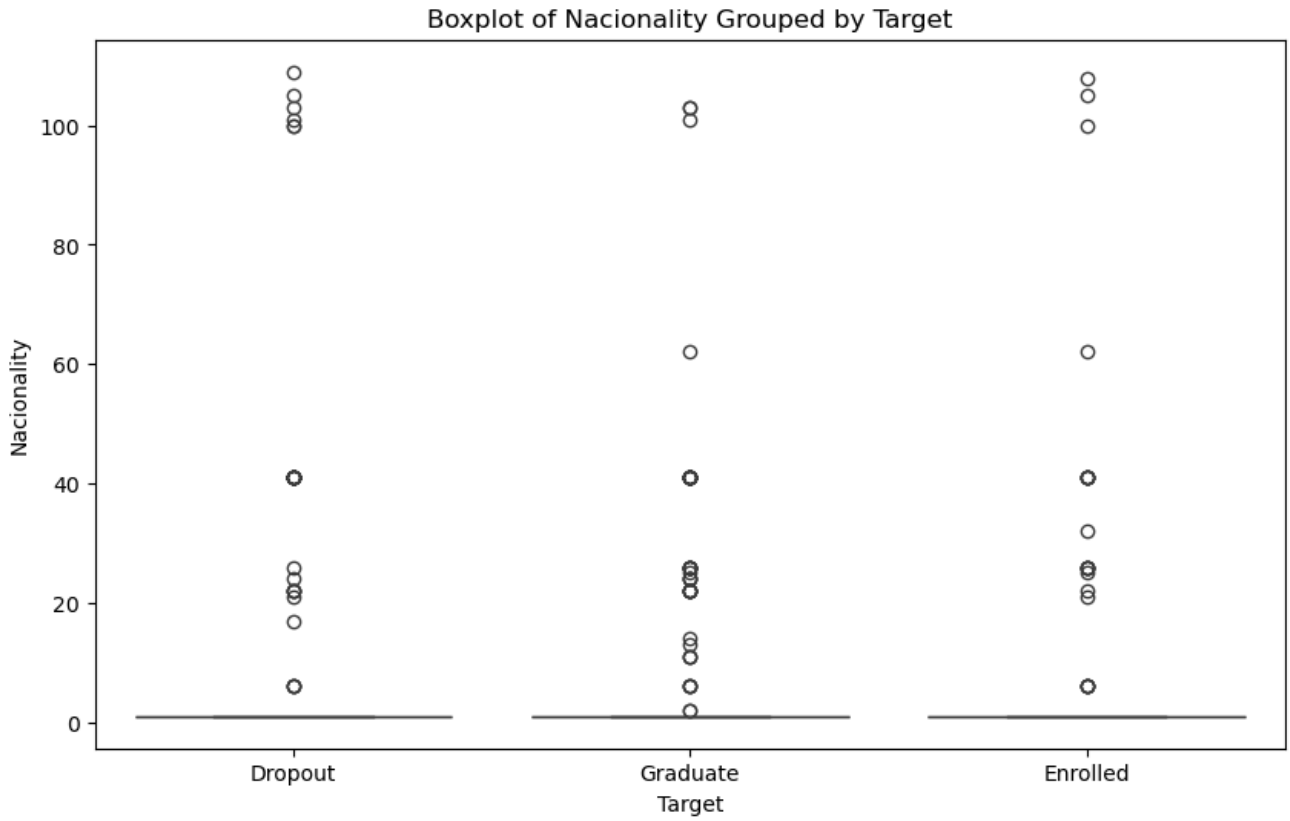


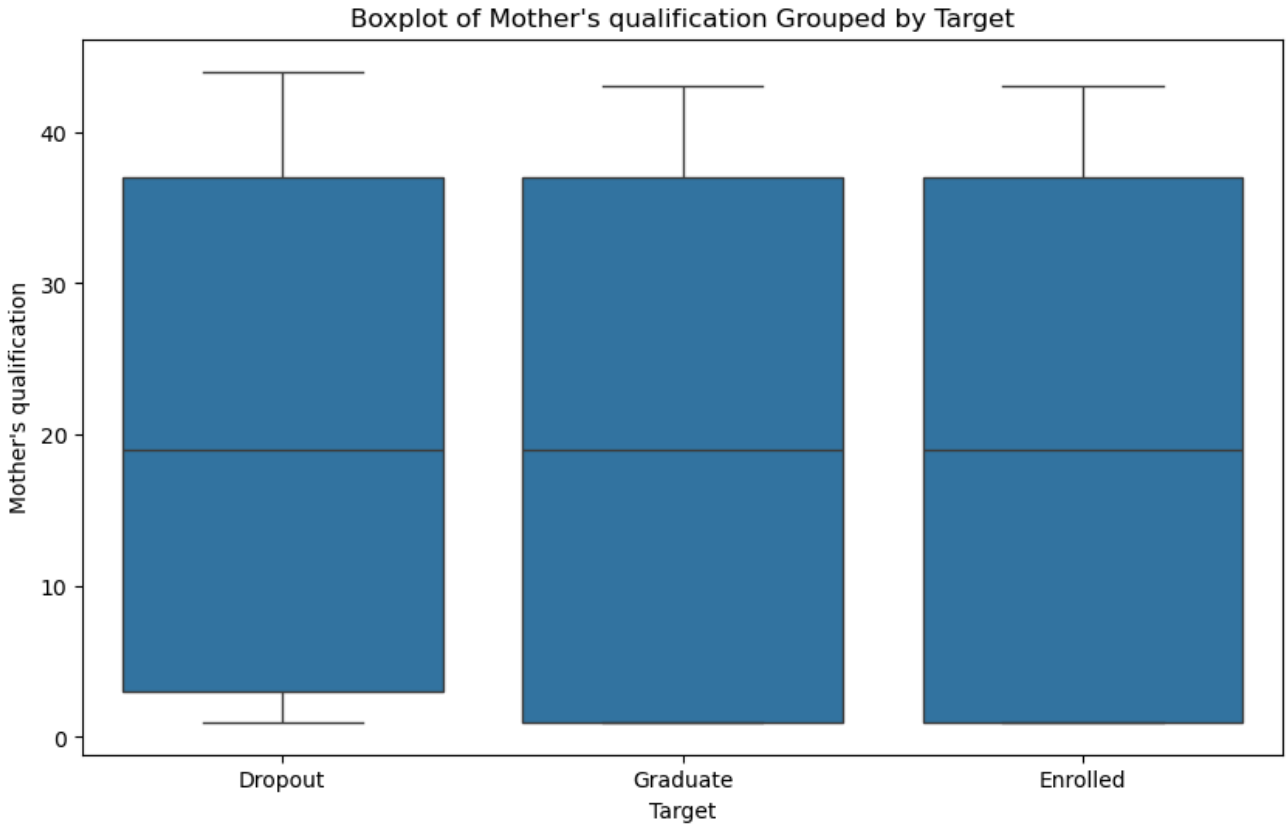

Daytime/
0.2
0.0
Dropout
Graduate
Enrolled
Target

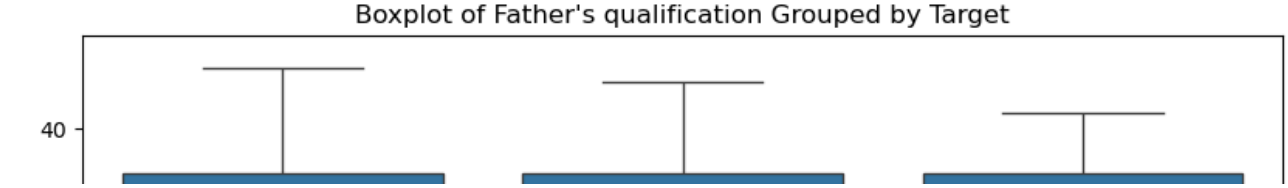

Boxplot of Previous qualification Grouped by Target
Previous qualification
40
30
20
10
0
Dropout
Graduate
Enrolled
Target

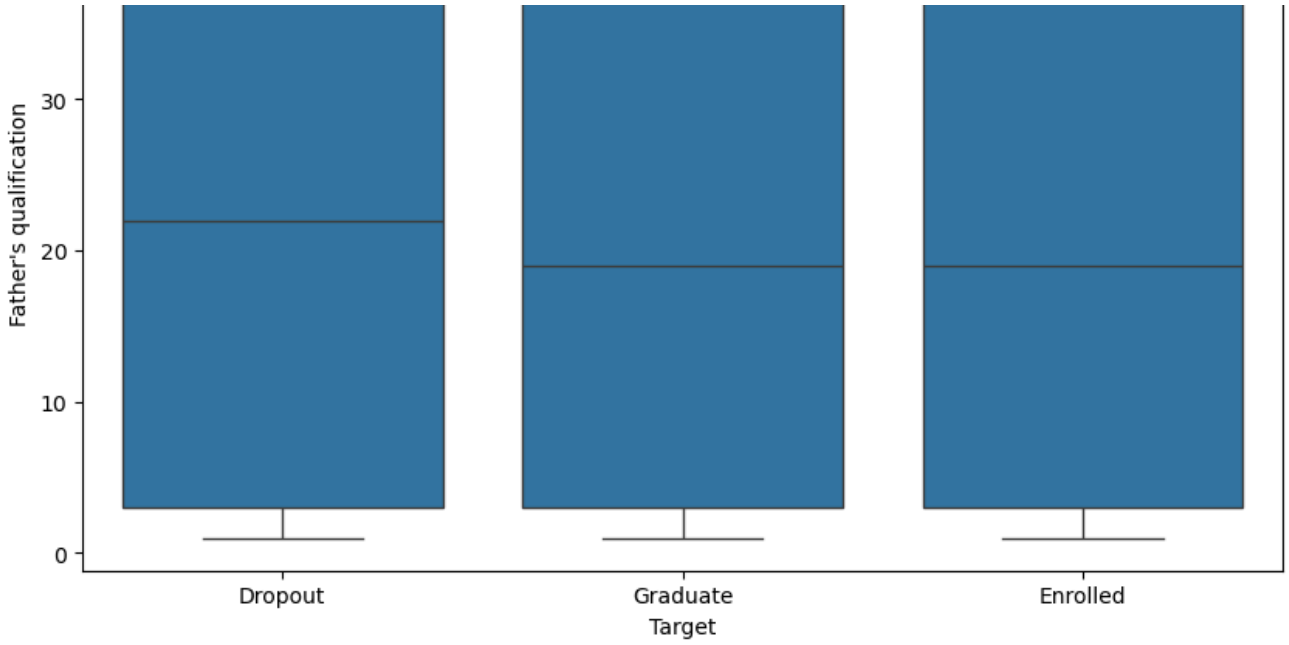

Boxplot of Previous qualification (grade) Grouped by Target
Previous qualification (grade)
180
160
140
120
100

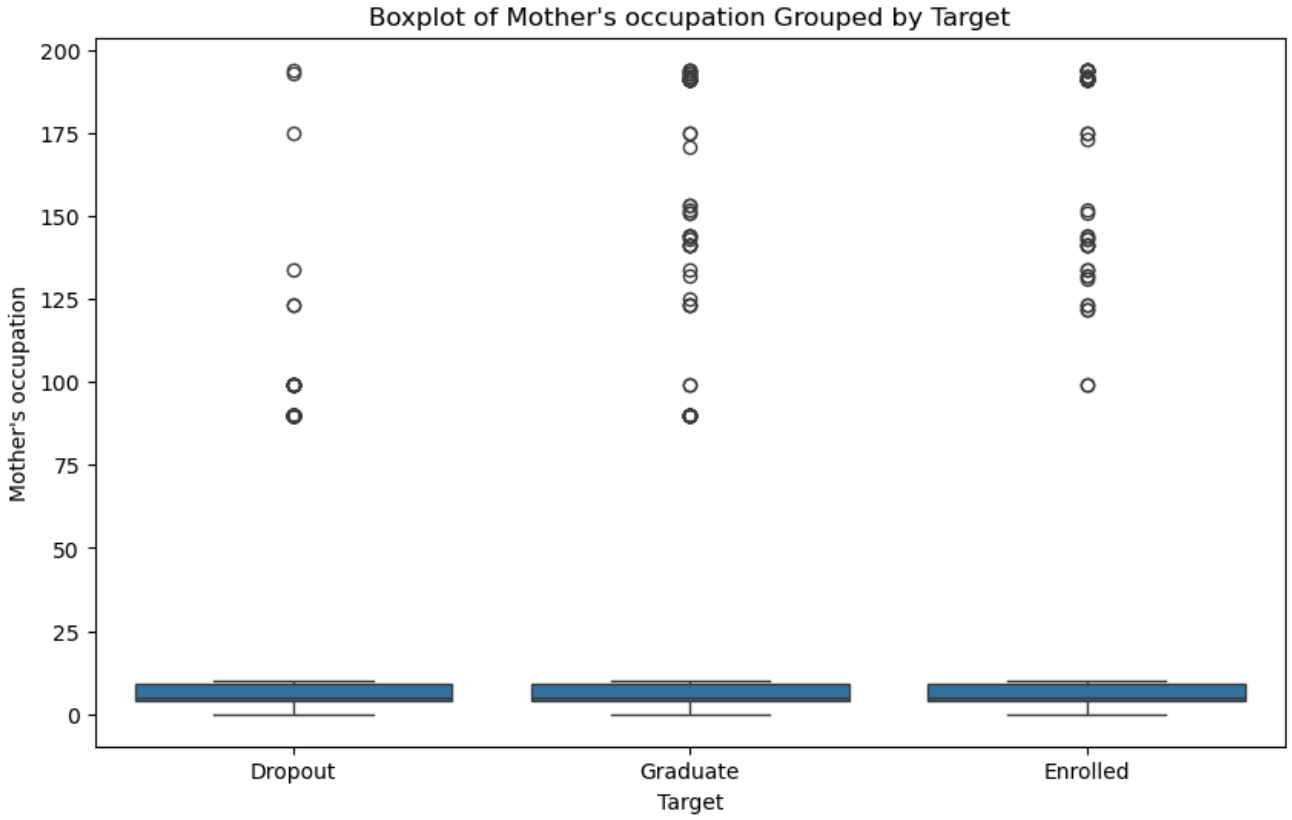

Dropout
Graduate
Enrolled
Target

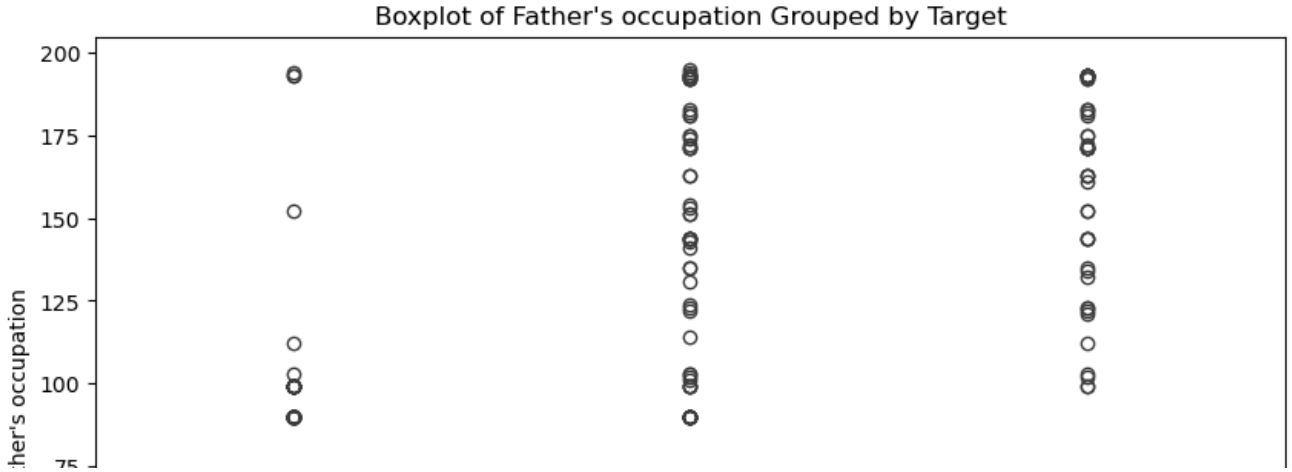

Boxplot of Nacionality Grouped by Target
Nacionality
100
80
60
40
20
0
Dropout
Graduate
Enrolled
Target

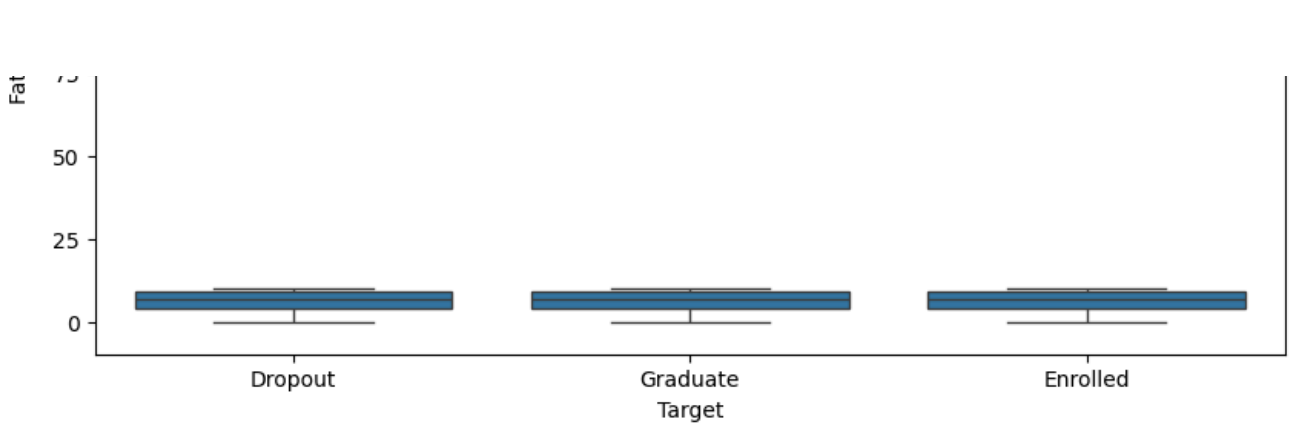

Boxplot of Mother's qualification Grouped by Target
Mother's qualification
40
30
20
10
0
Dropout
Graduate
Enrolled
Target

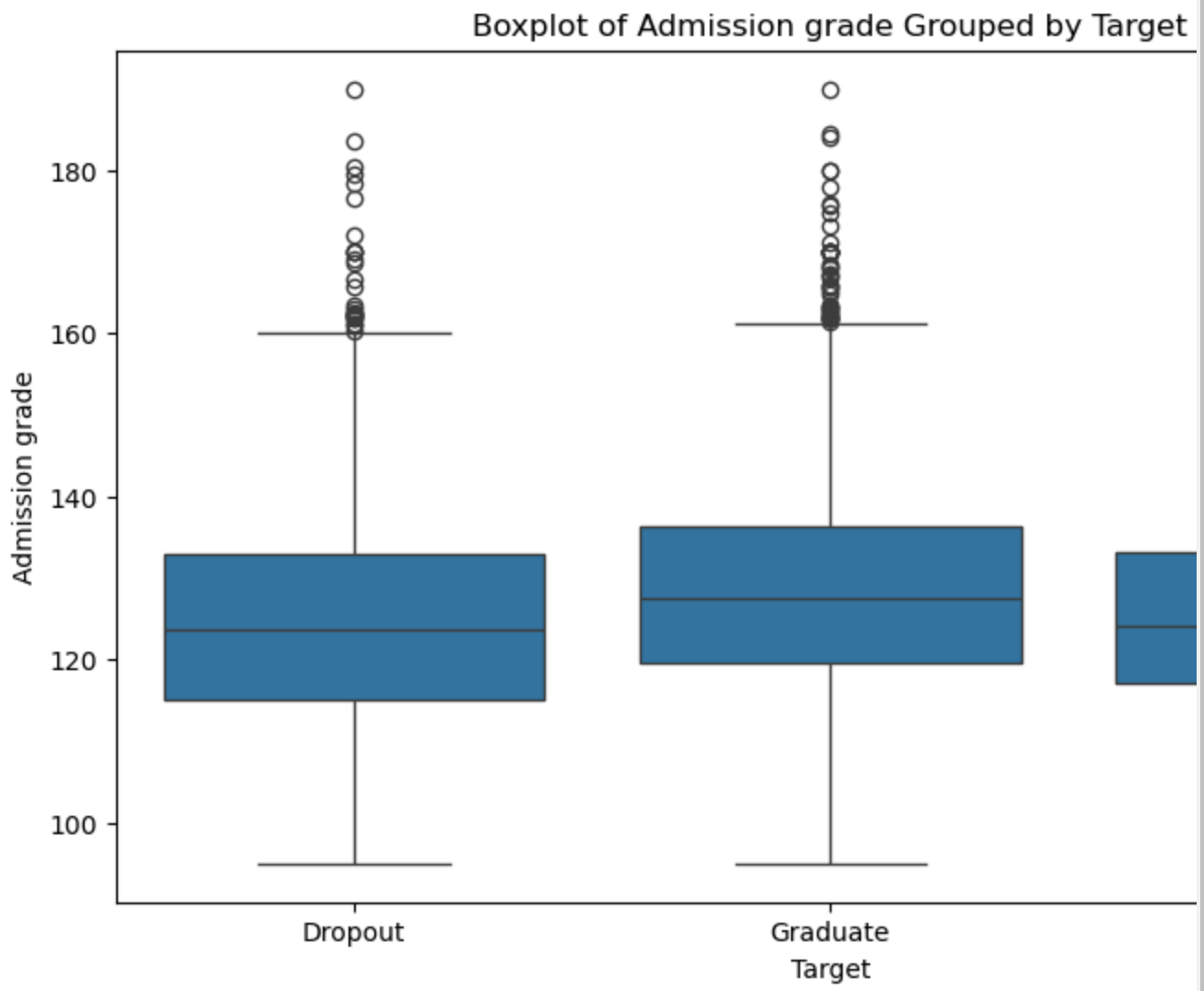

Boxplot of Father's qualification Grouped by Target
40

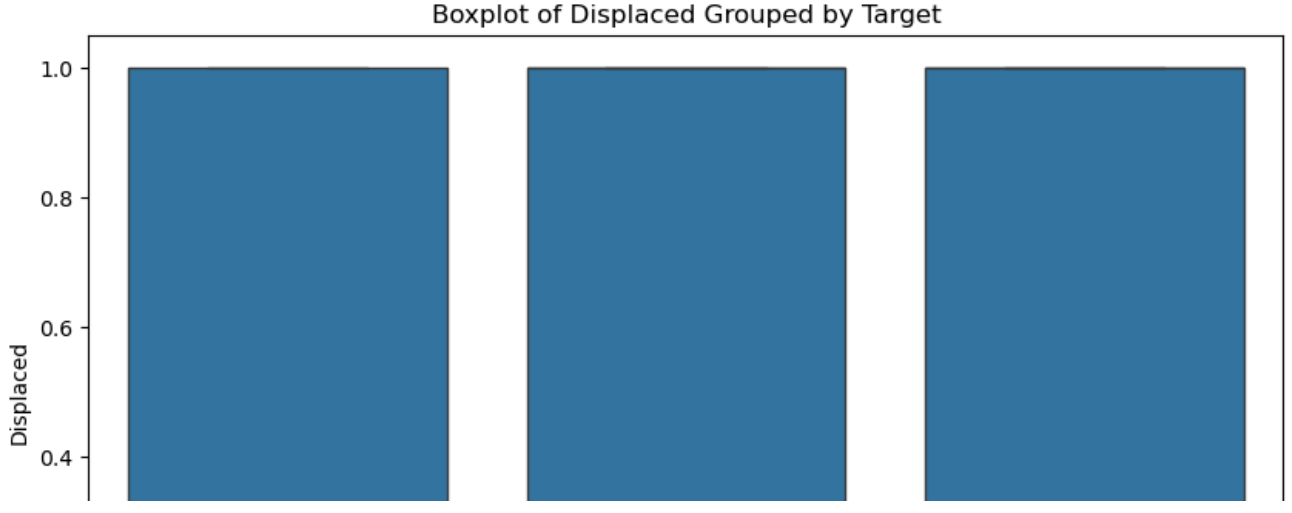

Father's qualification
30
20
10
0
Dropout
Graduate
Enrolled
Target

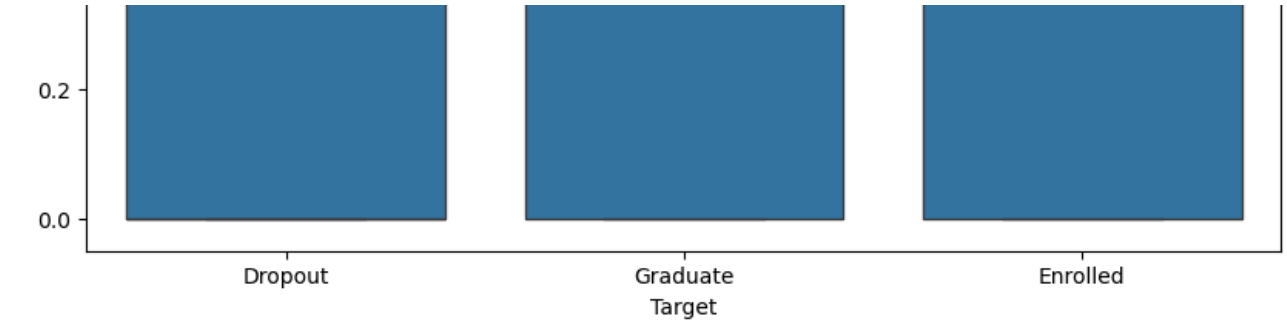

Boxplot of Mother's occupation Grouped by Target
Mother's occupation
200
175
150
125
100
75
50
25
0
Dropout
Graduate
Enrolled
Target

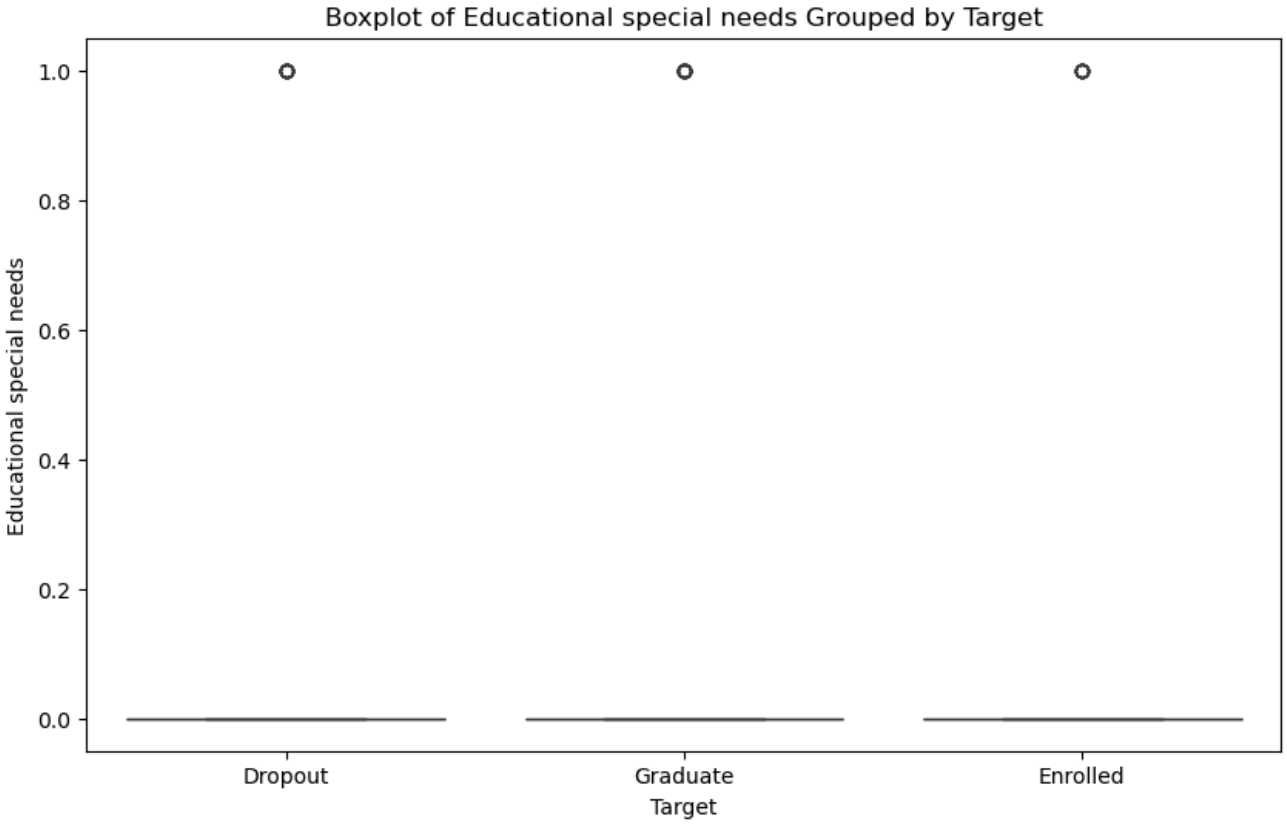

Boxplot of Father's occupation Grouped by Target
200
175
150
125
100

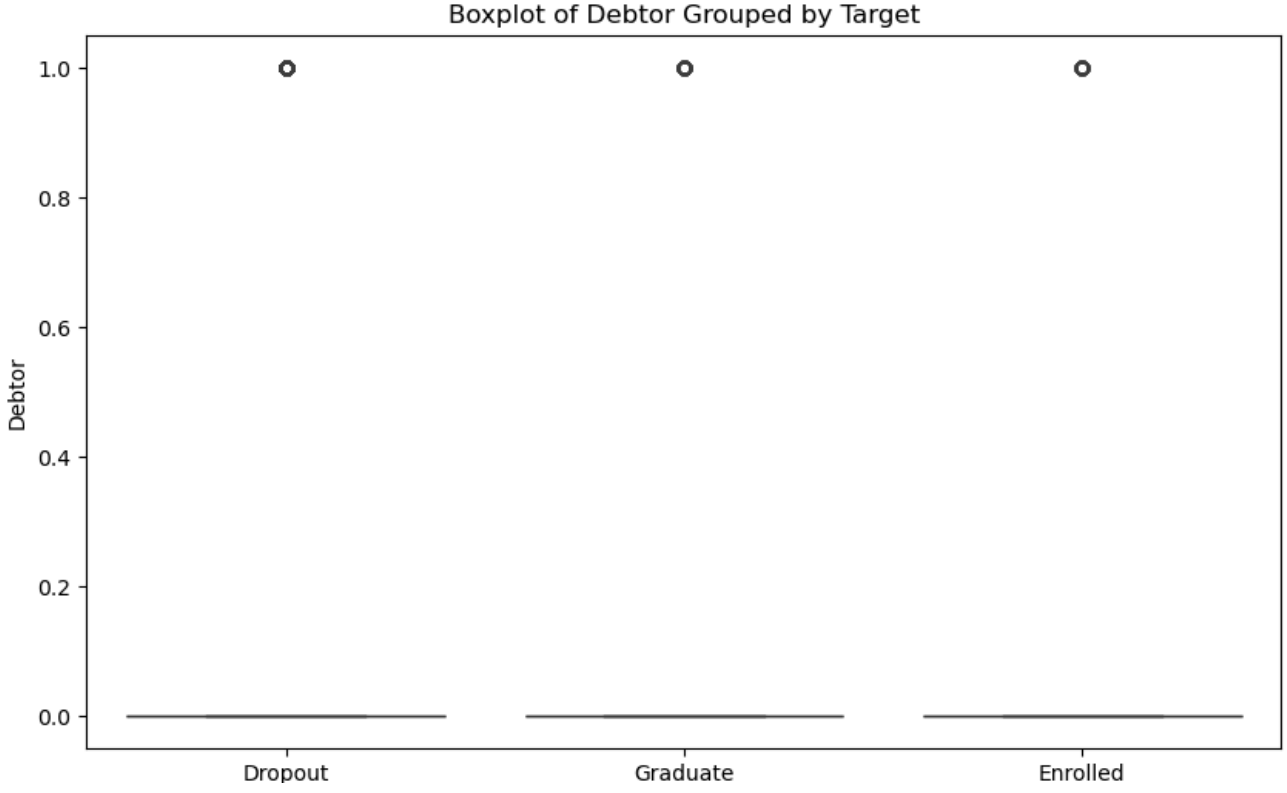

Fat
50
25
0
Dropout
Graduate
Enrolled
Target

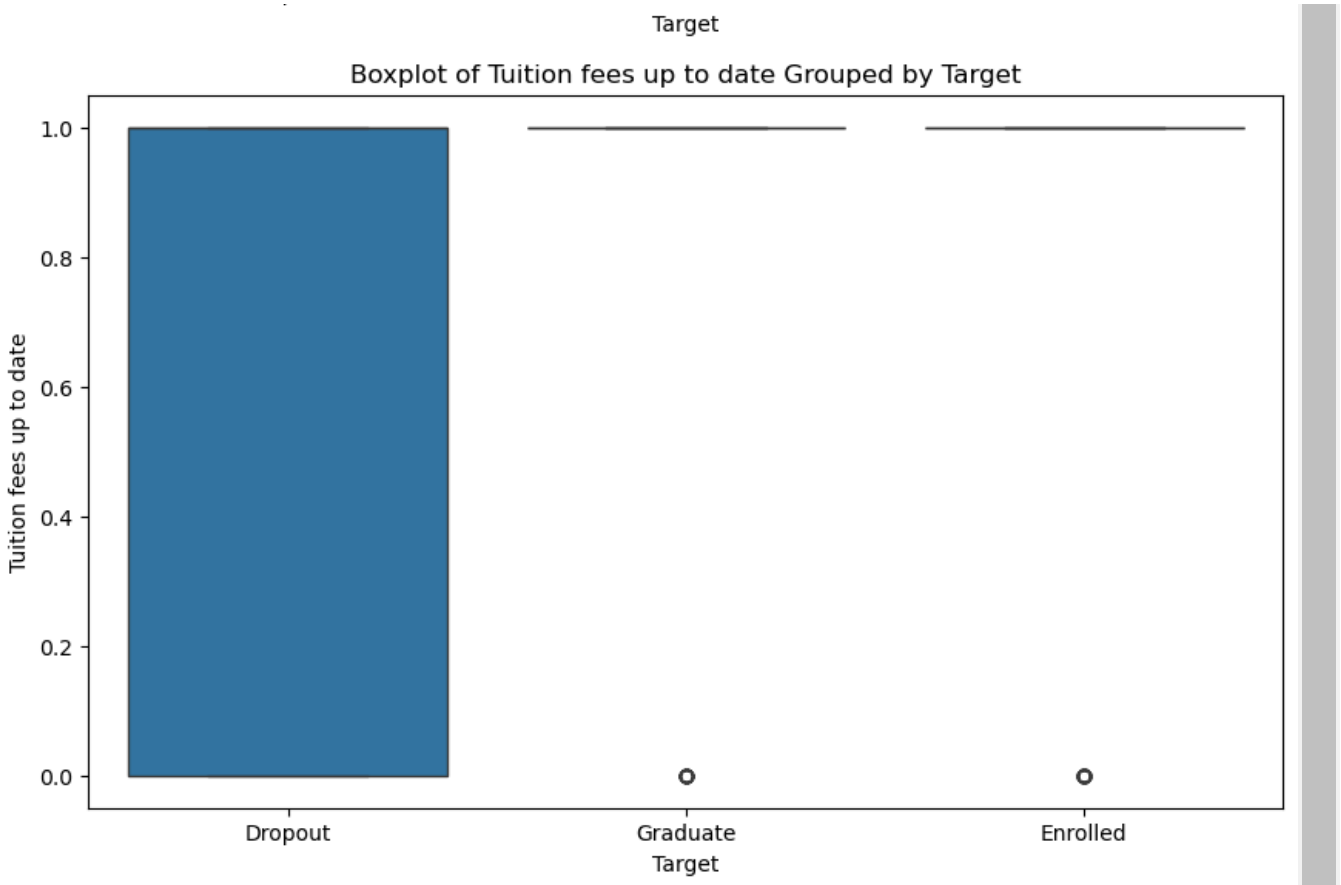

Boxplot of Admission grade Grouped by Target
Admission grade
180
160
140
120
100
Dropout
Graduate
Target


Boxplot of Displaced Grouped by Target
Displaced
1.0
0.8
0.6
0.4


0.2
0.0
Dropout
Graduate
Enrolled
Target

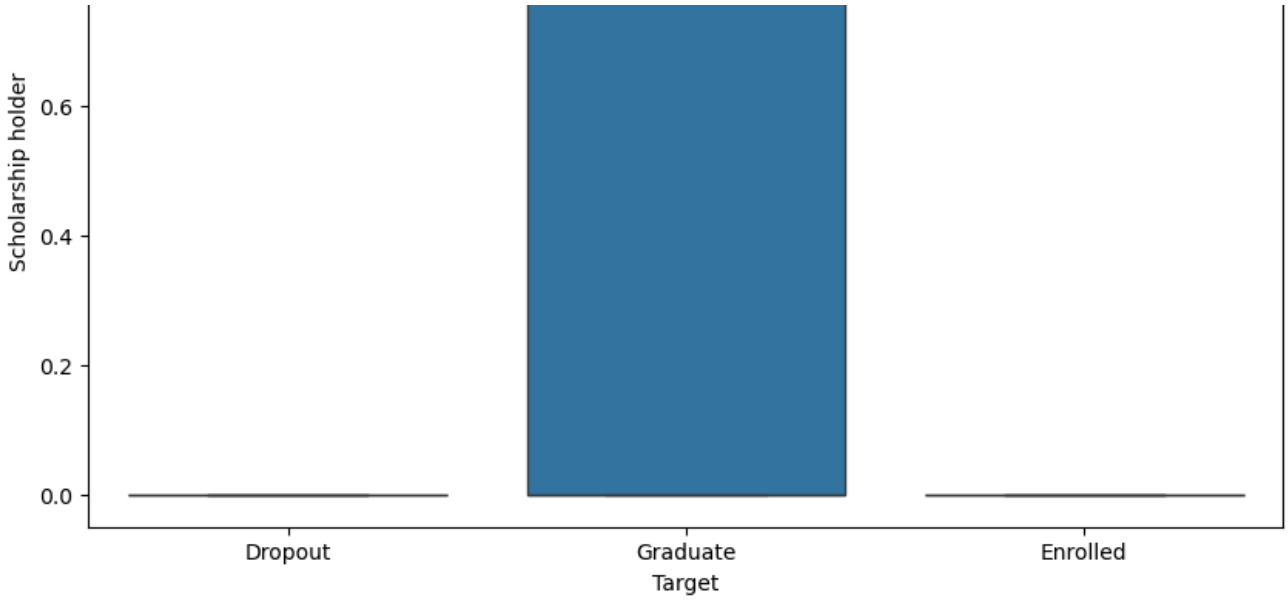

Boxplot of Educational special needs Grouped by Target
Educational special needs
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Graduate
Enrolled
Target

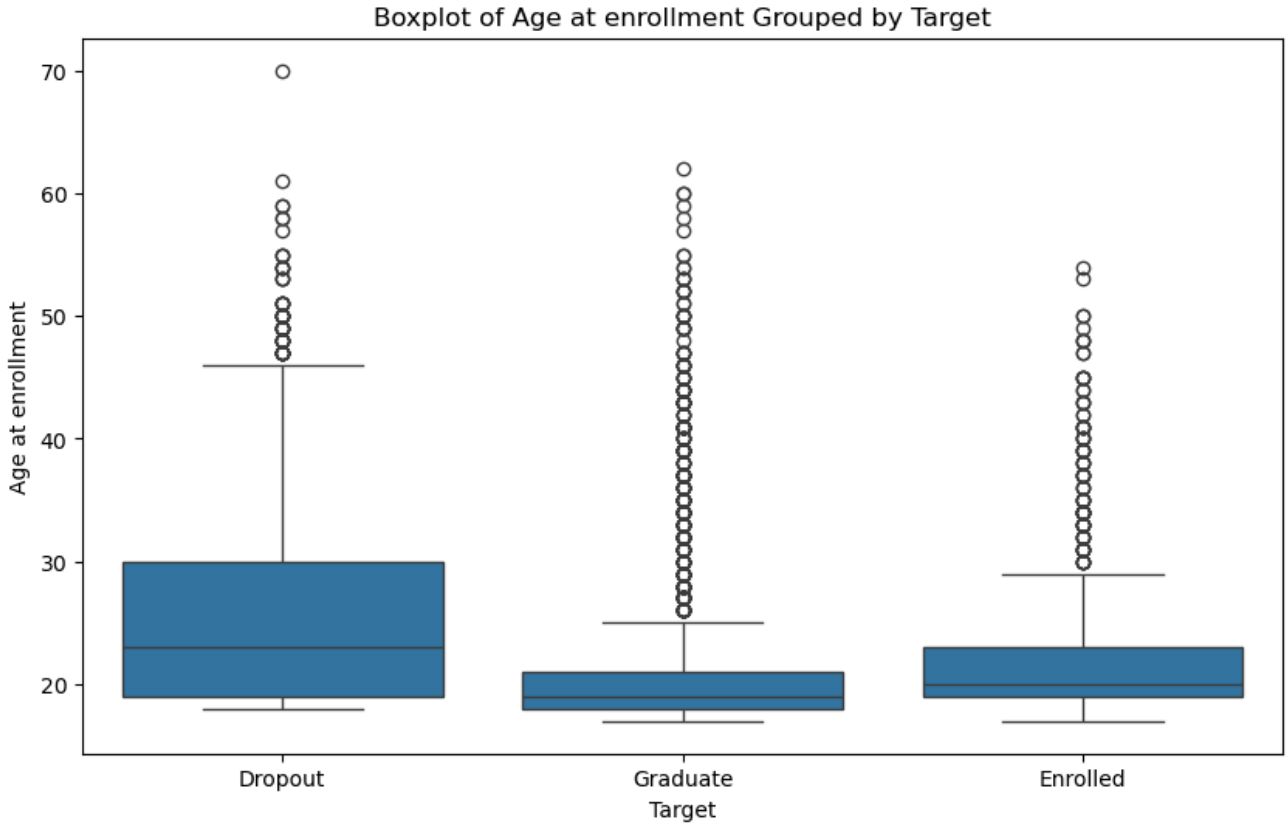

Boxplot of Debtor Grouped by Target
Debtor
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Graduate
Enrolled


Target
Boxplot of Tuition fees up to date Grouped by Target
Tuition fees up to date
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Graduate
Enrolled
Target

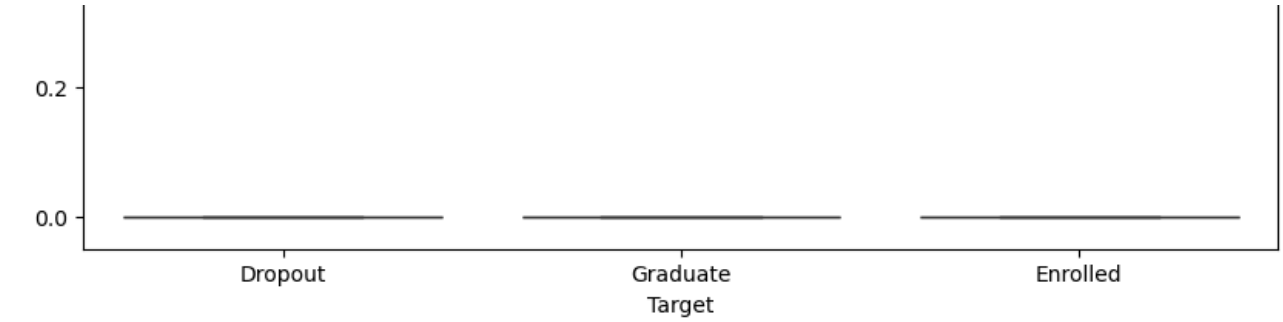

Boxplot of Gender Grouped by Target
Gender
1.0
0.8
0.6
0.4
0.2
0.0
Dropout
Graduate
Enrolled
Target

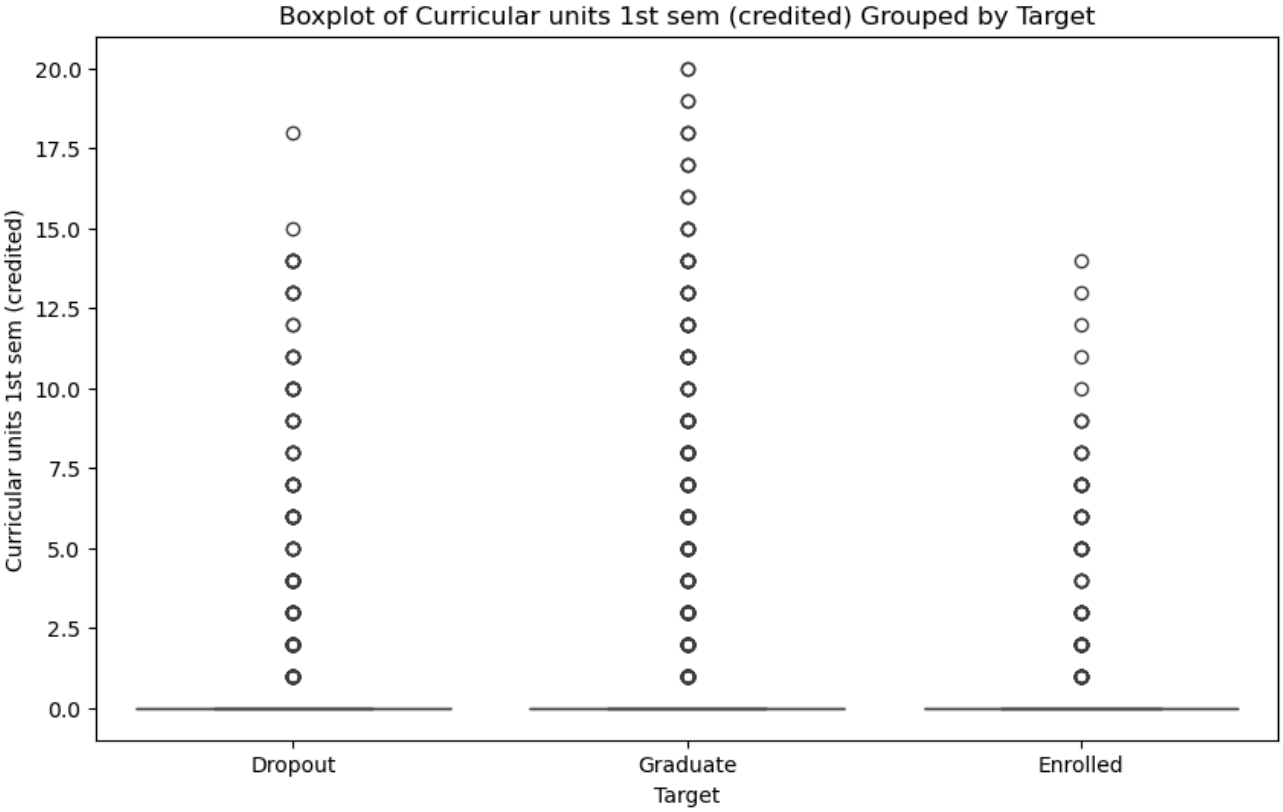

Boxplot of Scholarship holder Grouped by Target
1.0
0.8

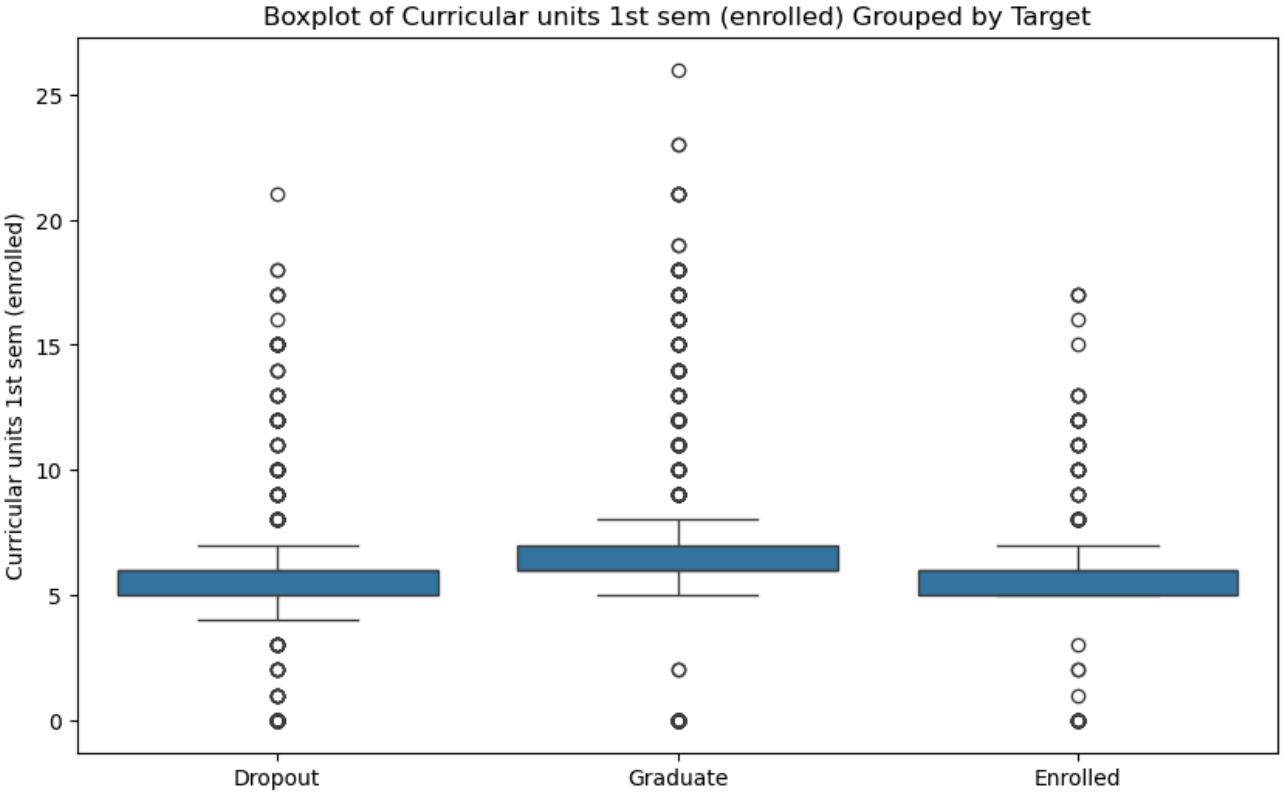

Scholarship holder
0.0
0.2
0.4
0.6
Dropout
Graduate
Enrolled
Target

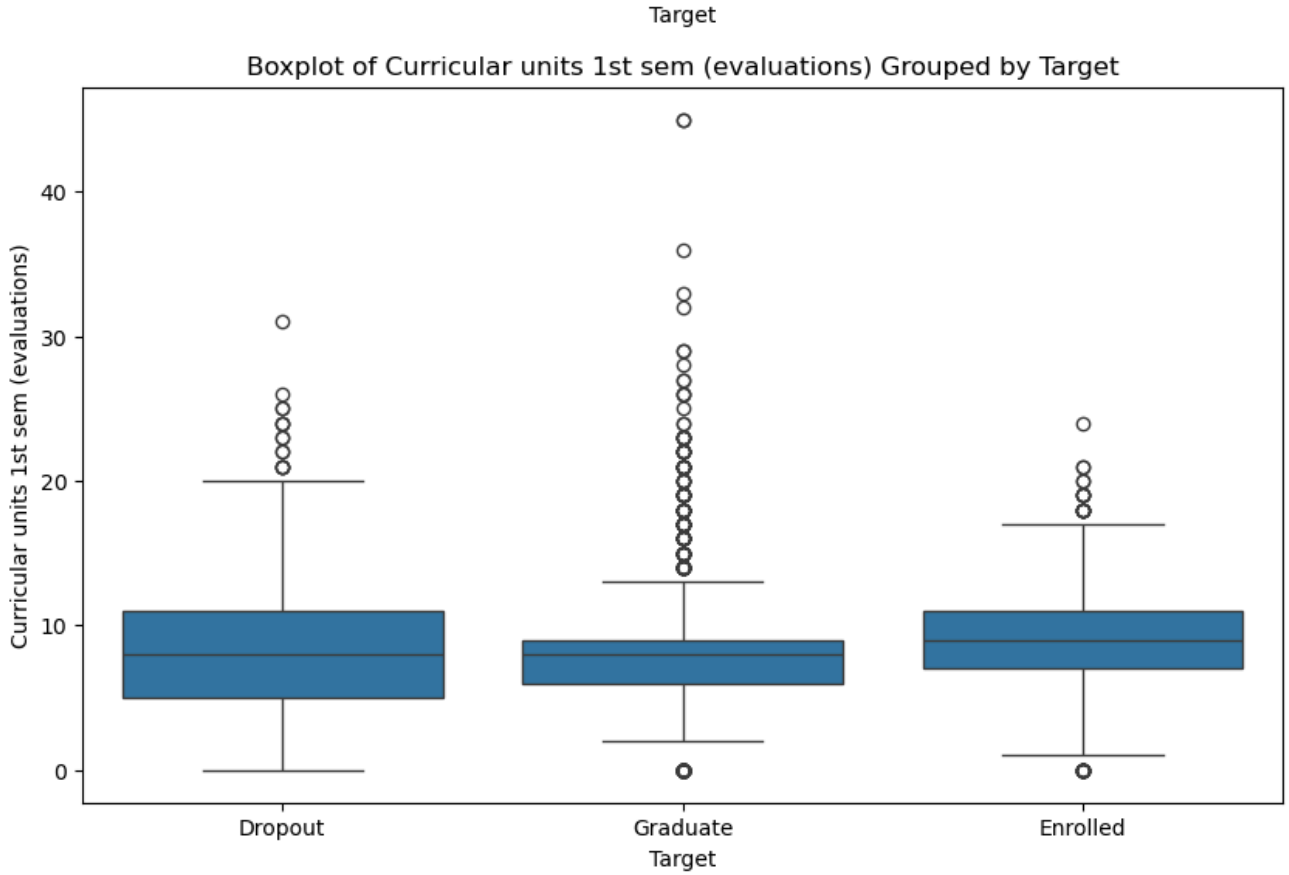

Boxplot of Age at enrollment Grouped by Target
Age at enrollment
20
30
40
50
60
70
Dropout
Graduate
Enrolled
Target

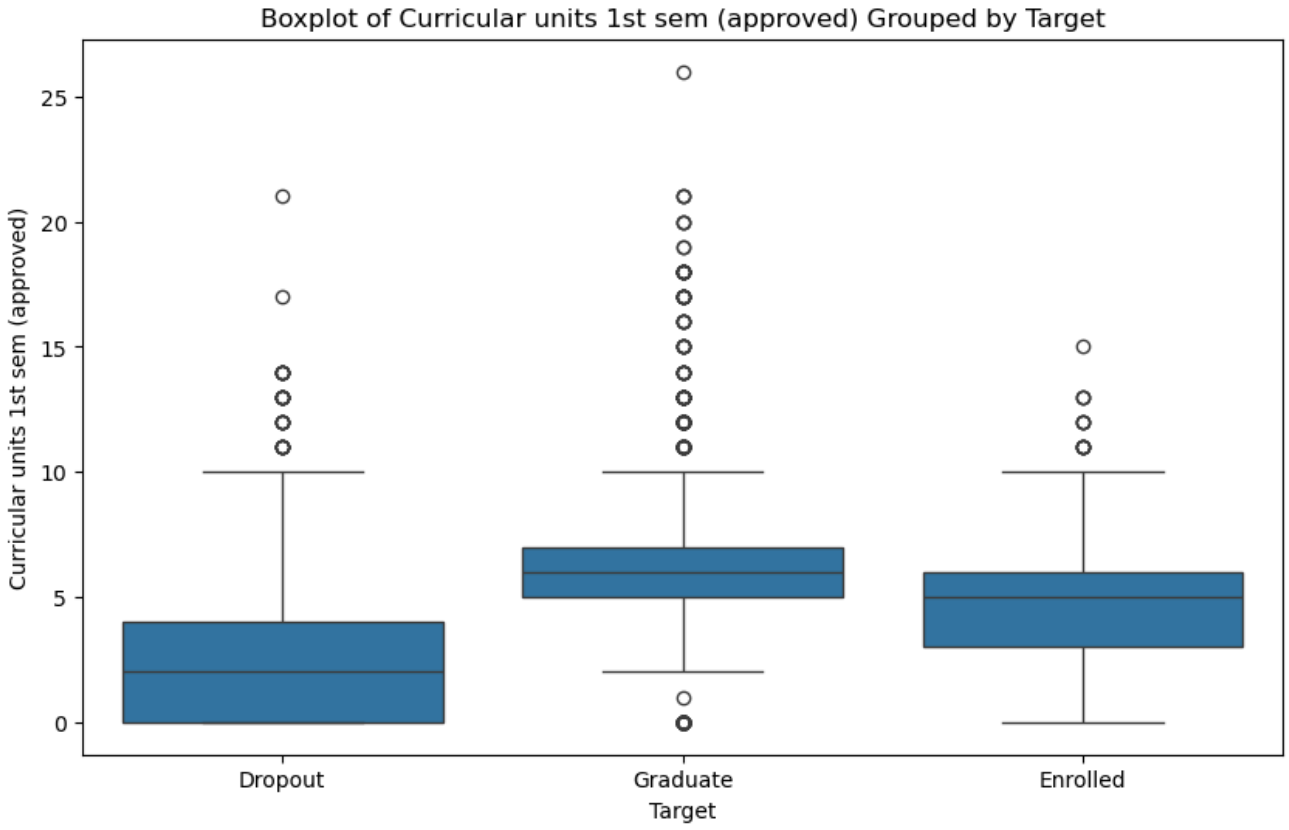

Boxplot of International Grouped by Target
International
0.4
0.6
0.8
1.0

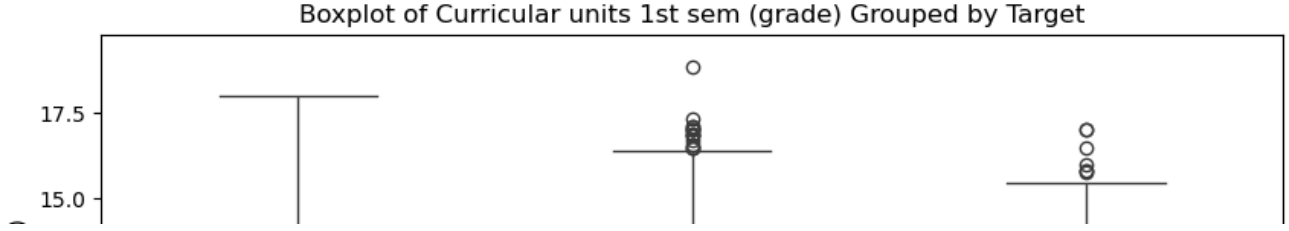

0.2
0.0
Dropout
Graduate
Enrolled
Target

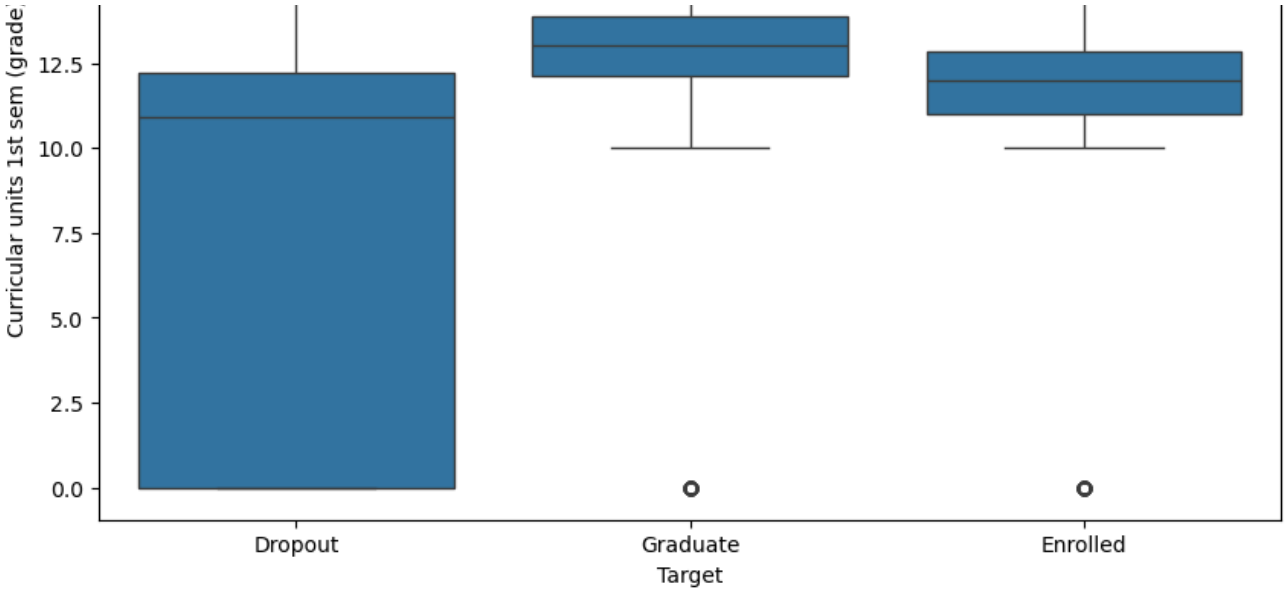

Boxplot of Curricular units 1st sem (credited) Grouped by Target
Curricular units 1st sem (credited)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Dropout
Graduate
Enrolled
Target

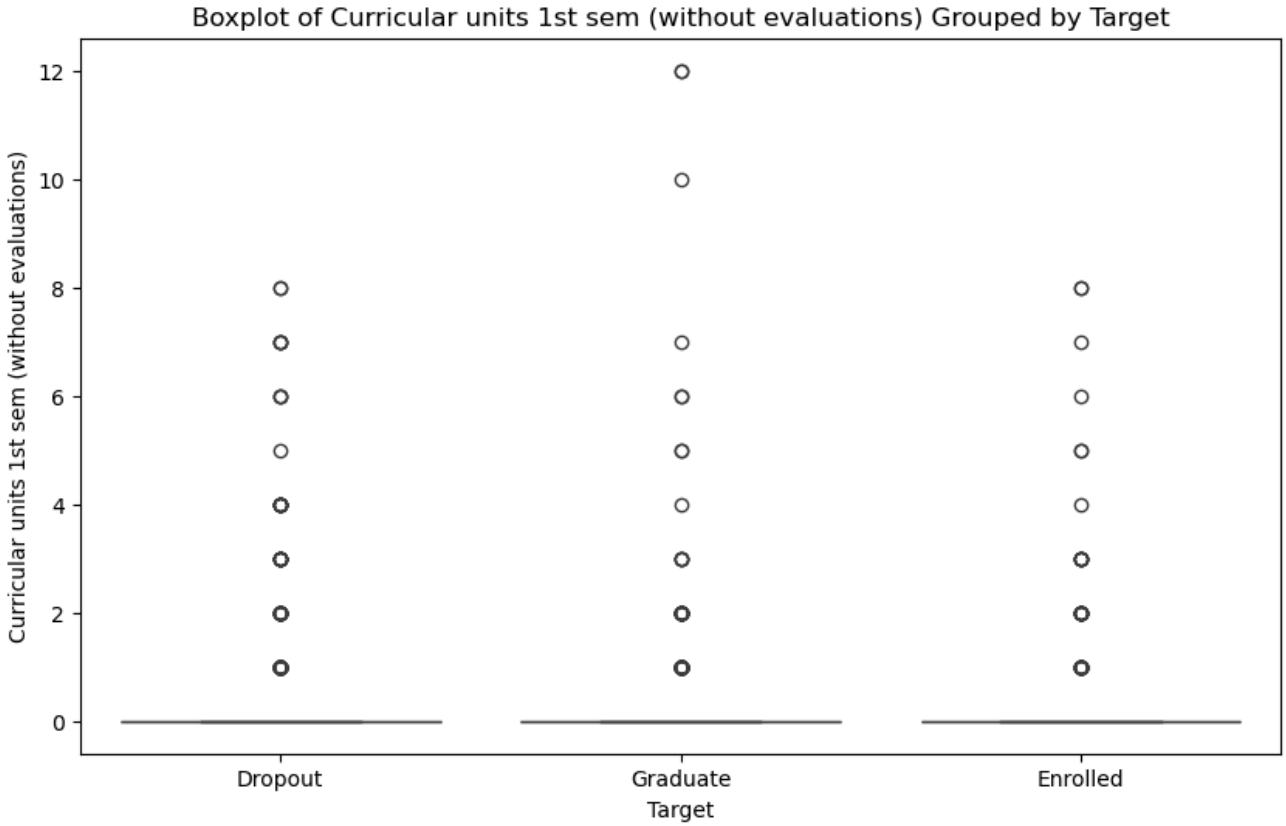

Boxplot of Curricular units 1st sem (enrolled) Grouped by Target
Curricular units 1st sem (enrolled)
25
20
15
10
5
0
Dropout
Graduate
Enrolled

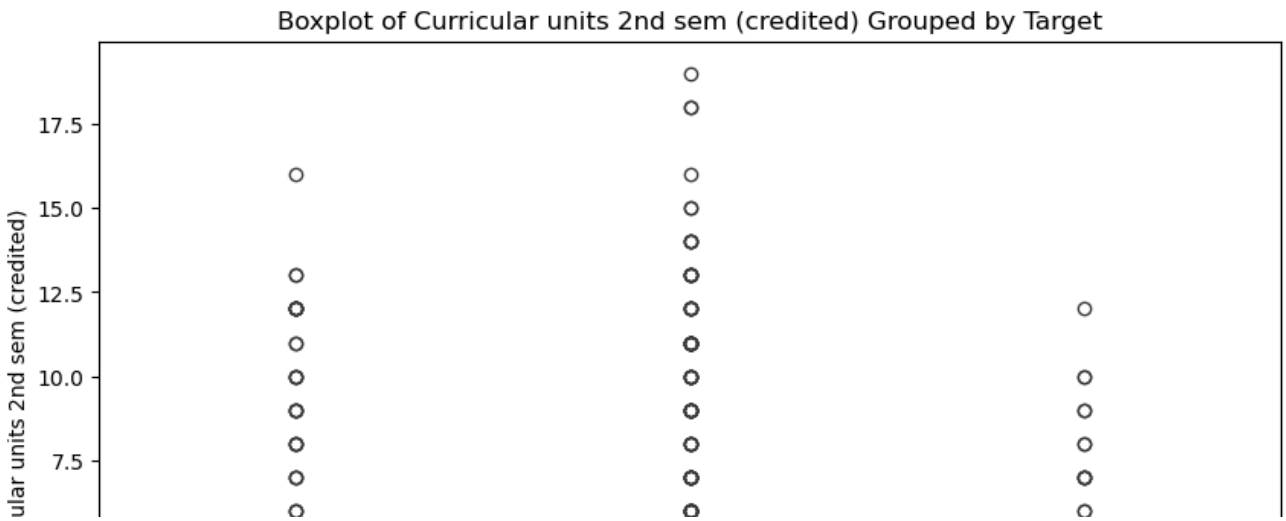

Target
Boxplot of Curricular units 1st sem (evaluations) Grouped by Target
Curricular units 1st sem (evaluations)
0
10
20
30
40
Dropout
Graduate
Enrolled
Target

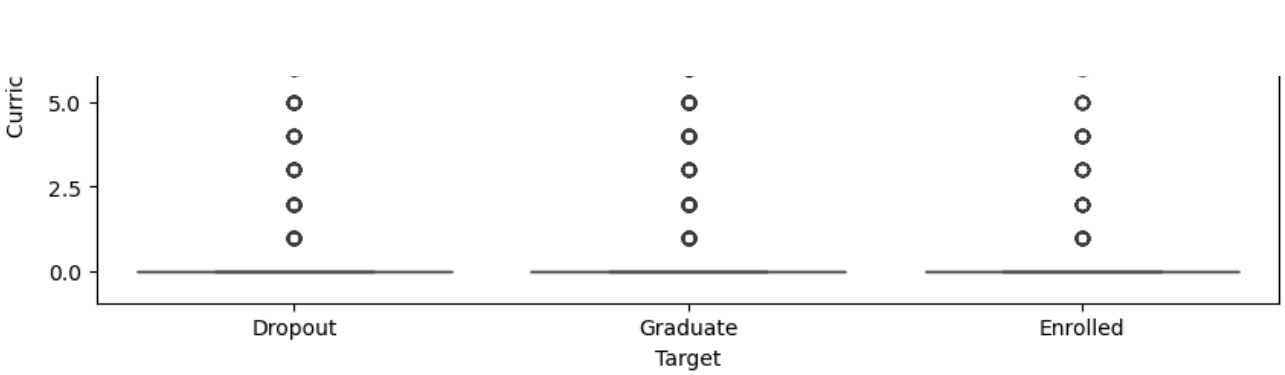

Boxplot of Curricular units 1st sem (approved) Grouped by Target
Curricular units 1st sem (approved)
0
5
10
15
20
25
Dropout
Graduate
Enrolled
Target

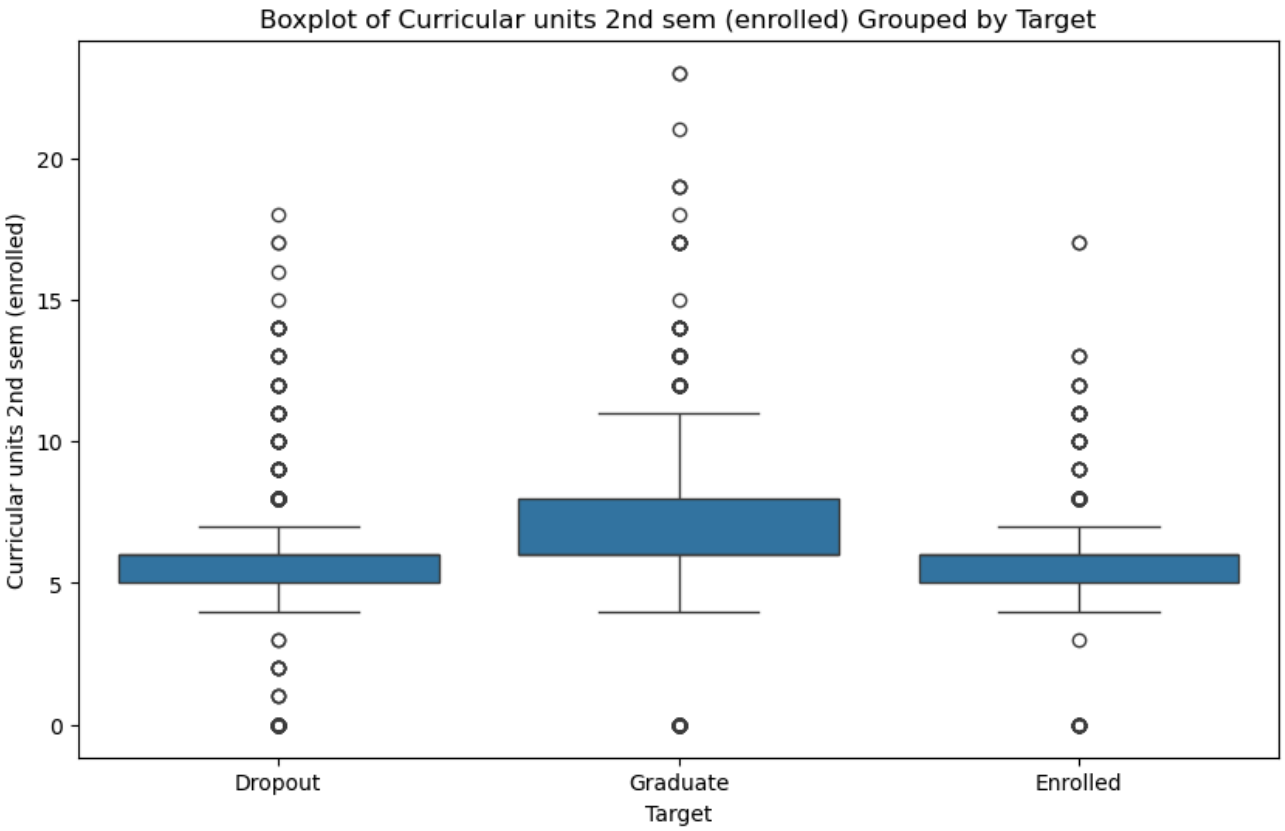

Boxplot of Curricular units 1st sem (grade) Grouped by Target
15.0
17.5

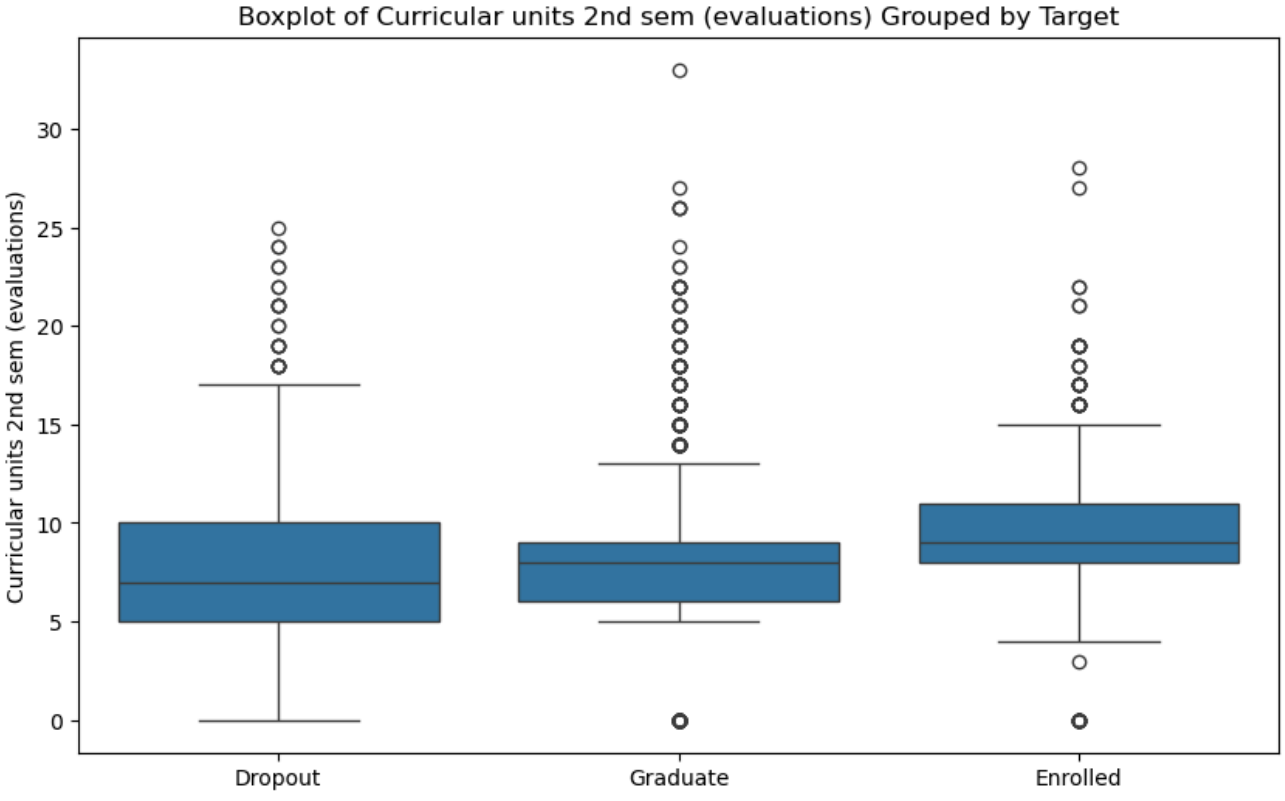

Curricular units 1st sem (grade)
12.5
10.0
7.5
5.0
2.5
0.0
Dropout
Graduate
Enrolled
Target

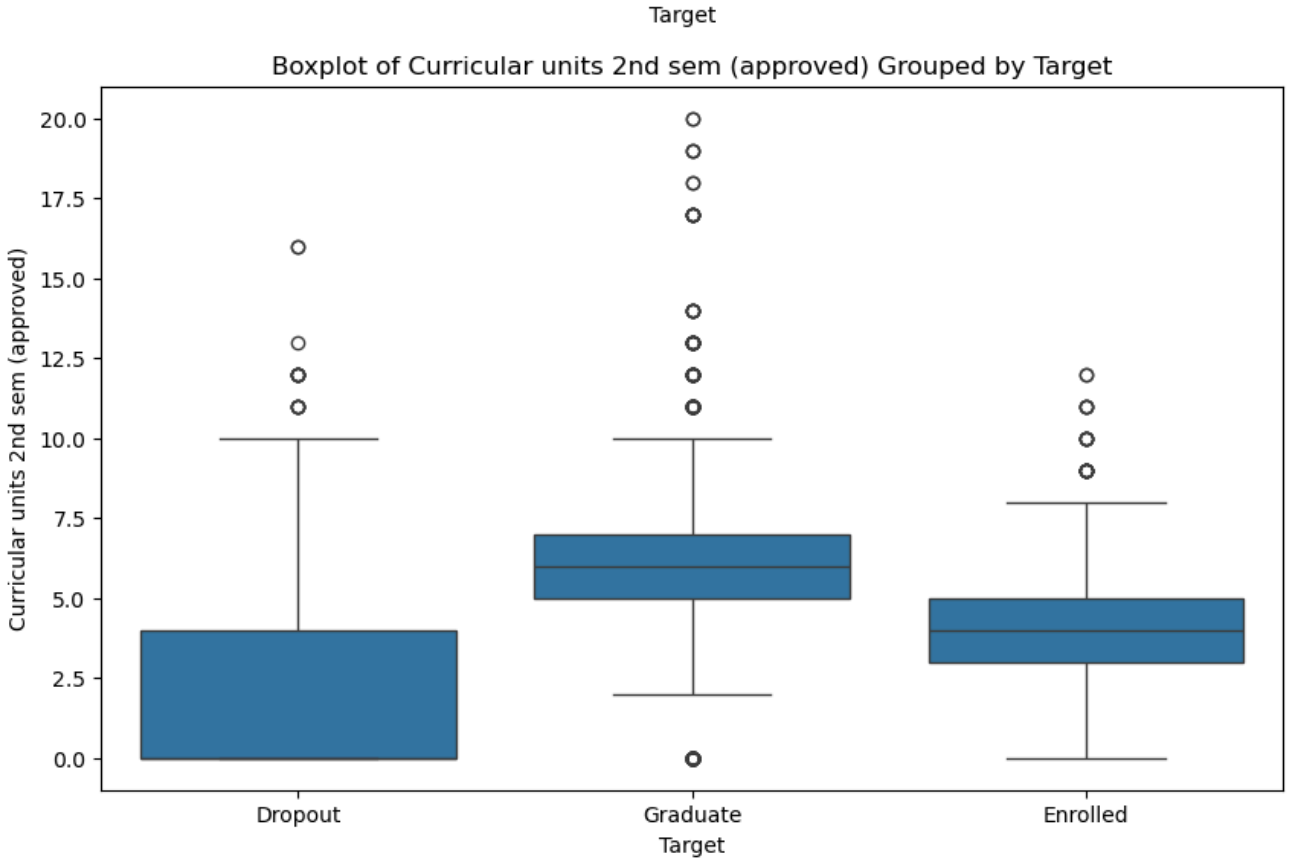

Boxplot of Curricular units 1st sem (without evaluations) Grouped by Target
Curricular units 1st sem (without evaluations)
12
10
8
6
4
2
0
Dropout
Graduate
Enrolled
Target

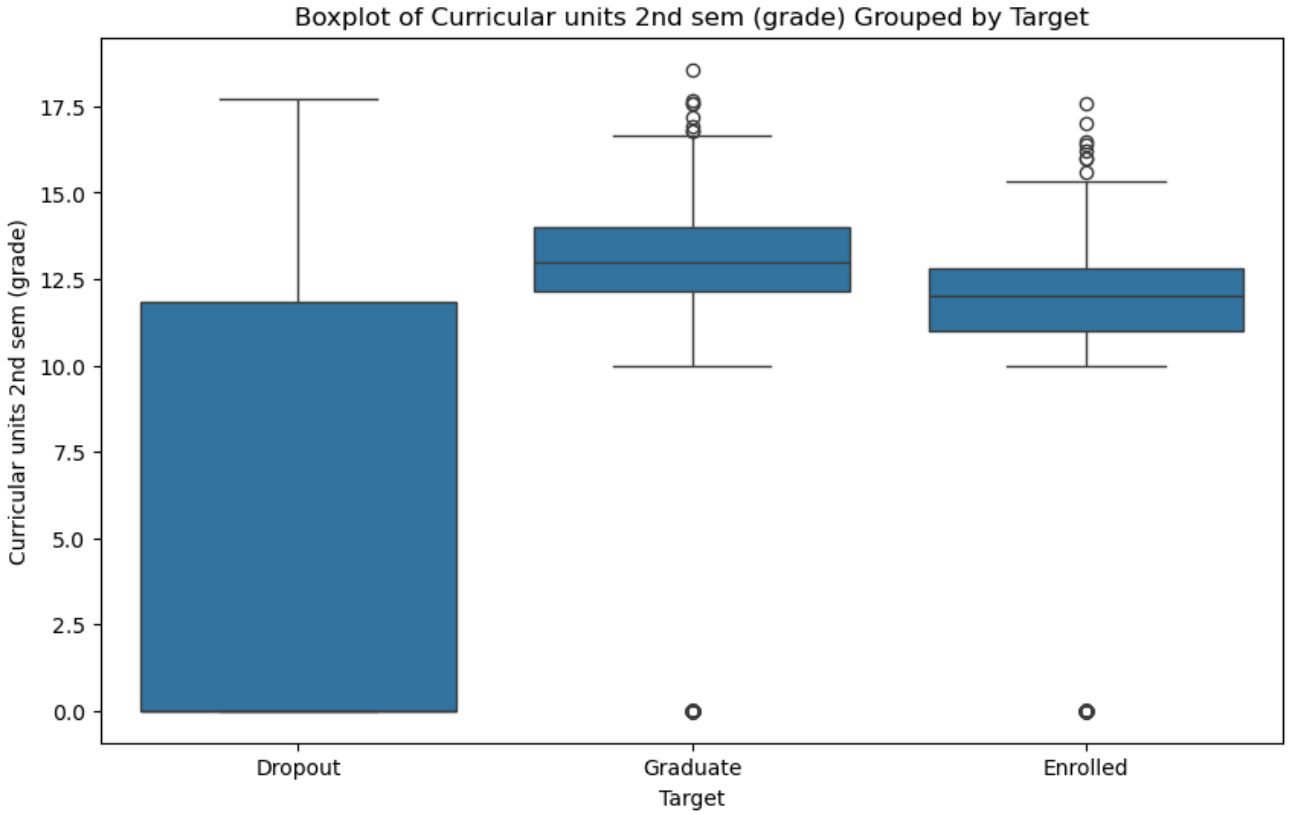

Boxplot of Curricular units 2nd sem (credited) Grouped by Target
ular units 2nd sem (credited)
17.5
15.0
12.5
10.0
7.5

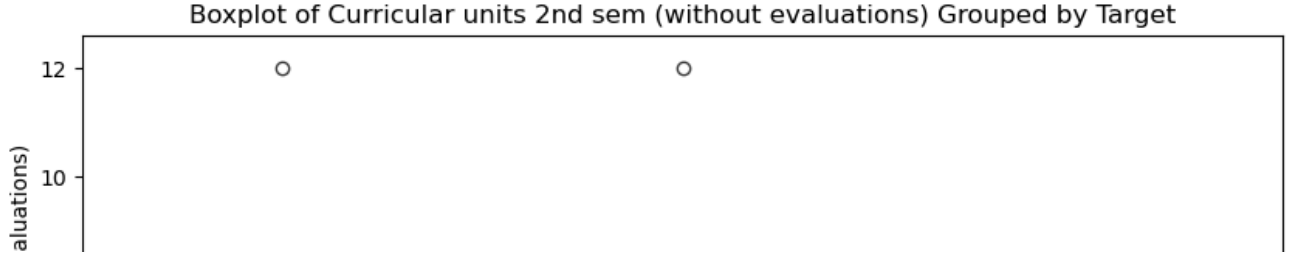

Curric
5.0
2.5
0.0
Dropout
Graduate
Enrolled
Target

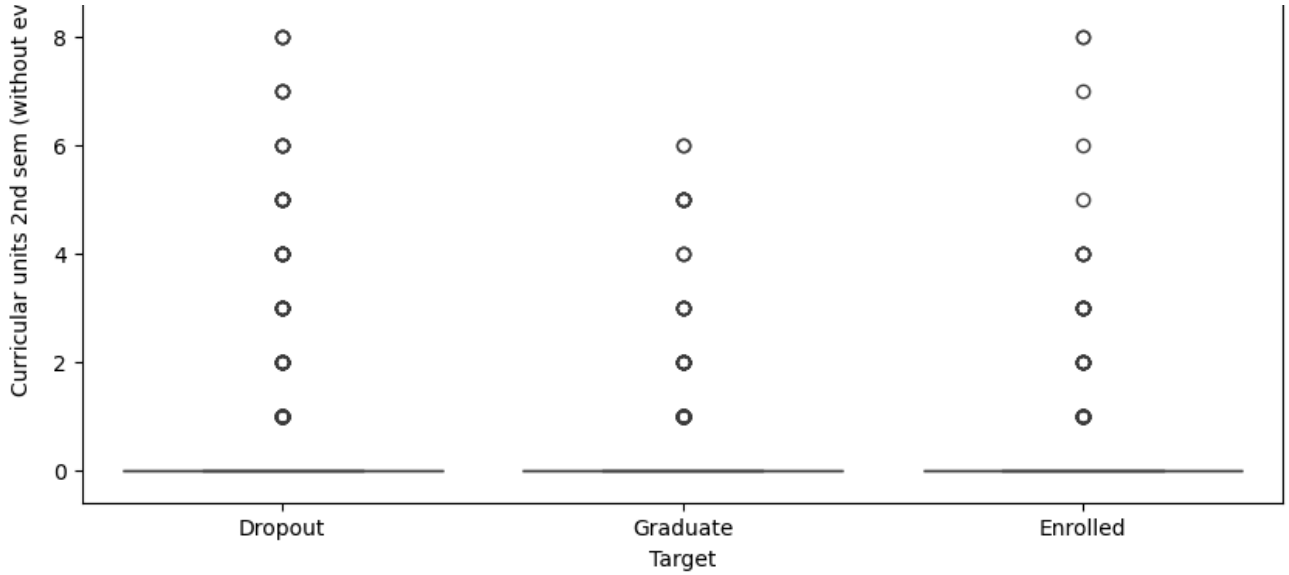

Boxplot of Curricular units 2nd sem (enrolled) Grouped by Target
Curricular units 2nd sem (enrolled)
20
15
10
5
0
Dropout
Graduate
Enrolled
Target

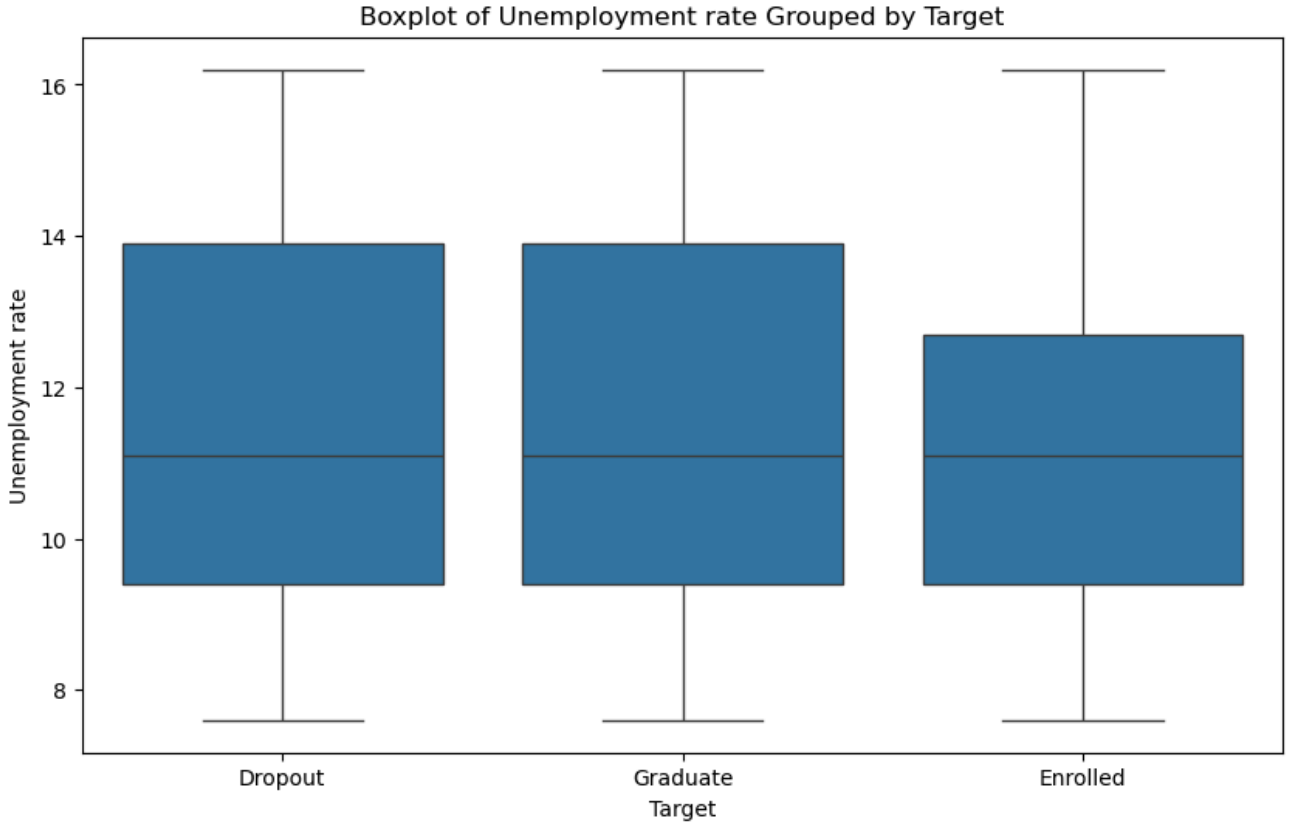

Boxplot of Curricular units 2nd sem (evaluations) Grouped by Target
Curricular units 2nd sem (evaluations)
30
25
20
15
10
5
0
Dropout
Graduate
Enrolled


Target
Boxplot of Curricular units 2nd sem (approved) Grouped by Target
Curricular units 2nd sem (approved)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Dropout
Graduate
Enrolled
Target

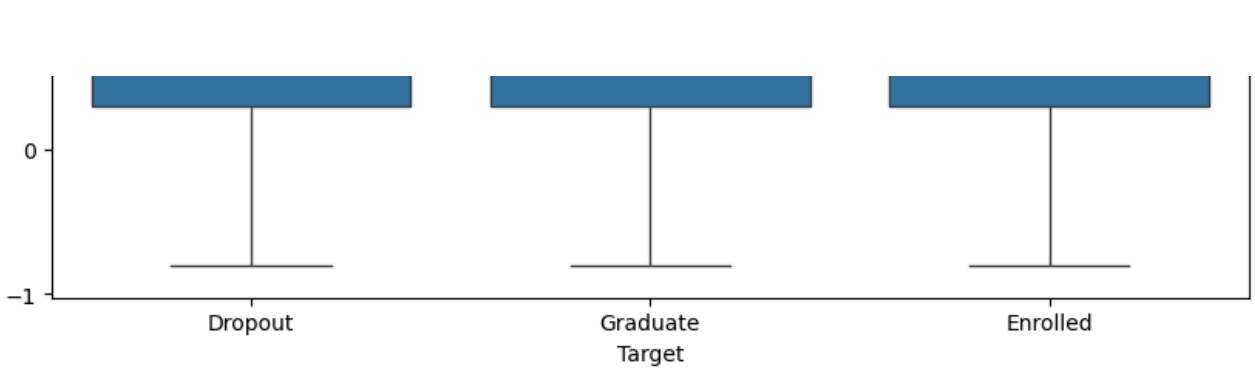

Boxplot of Curricular units 2nd sem (grade) Grouped by Target
Curricular units 2nd sem (grade)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Dropout
Graduate
Enrolled
Target

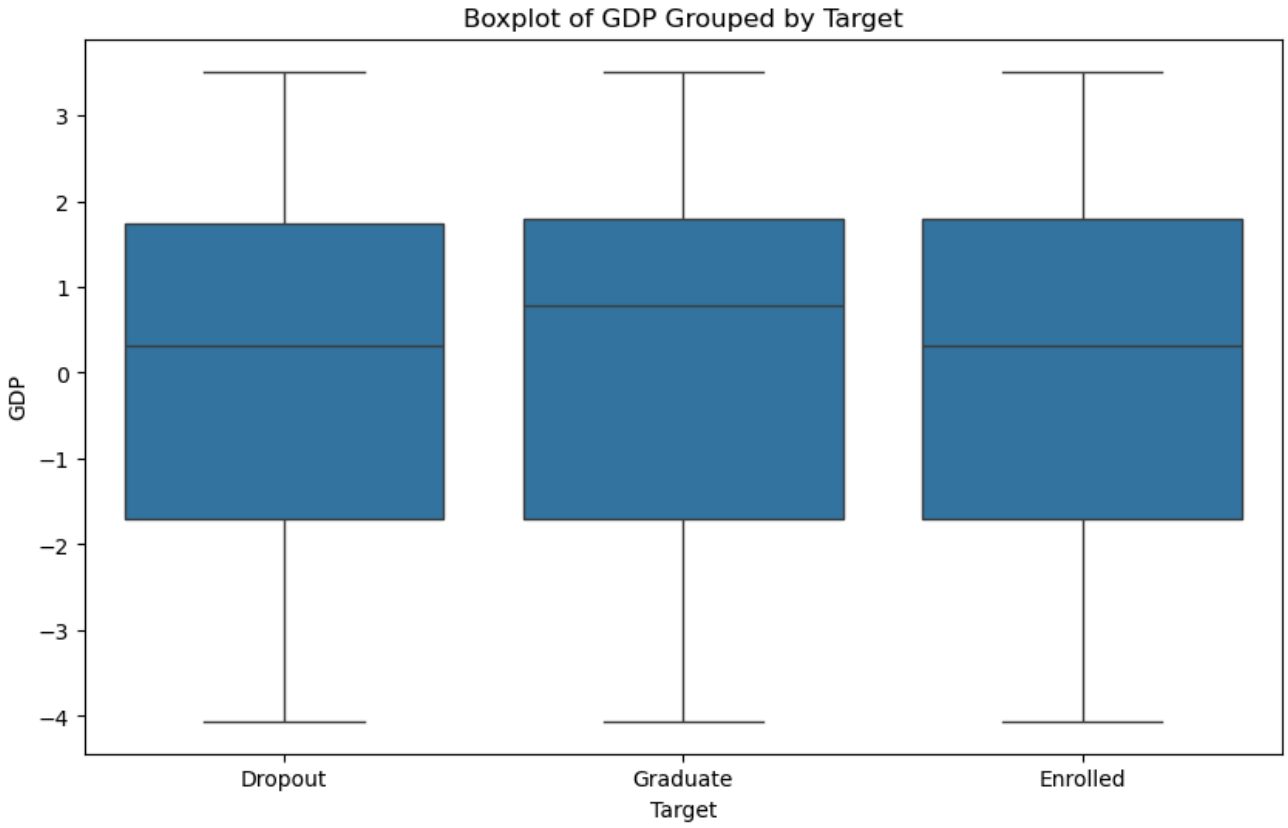

Boxplot of Curricular units 2nd sem (without evaluations) Grouped by Target
12
10
aluations)


Curricular units 2nd sem (without ev
8
6
4
2
0
Dropout
Graduate
Enrolled
Target


Boxplot of Unemployment rate Grouped by Target
Unemployment rate
16
14
12
10
8
Dropout
Graduate
Enrolled
Target


Boxplot of Inflation rate Grouped by Target
Inflation rate
3
2
1


0
−1
Dropout
Graduate
Enrolled
Target


Boxplot of GDP Grouped by Target
GDP
3
2
1
0
−1
−2
−3
−4
Dropout
Graduate
Enrolled
Target