

Dataset Overview

Introduction

The dataset contains information related to students' demographics, academic performance, and various socio-economic factors. The goal is to predict students' dropout and academic success based on the provided features.

Dataset Columns

1. **Marital status:** Marital status of the student.
2. **Application mode:** Mode of application for enrollment.
3. **Application order:** Order in which the application was received.
4. **Course:** Course enrolled in by the student.
5. **Daytime/evening attendance:** Type of attendance (daytime or evening).
6. **Previous qualification:** Student's previous educational qualification.
7. **Nationality:** Nationality of the student.
8. **Mother's qualification:** Qualification of the student's mother.
9. **Father's qualification:** Qualification of the student's father.
10. **Mother's occupation:** Occupation of the student's mother.
11. **Father's occupation:** Occupation of the student's father.
12. **Displaced:** Whether the student is displaced.
13. **Educational special needs:** Whether the student has special educational needs.
14. **Debtor:** Whether the student has debts related to tuition fees.
15. **Tuition fees up to date:** Status of payment of tuition fees.
16. **Gender:** Gender of the student.
17. **Scholarship holder:** Whether the student holds a scholarship.
18. **Age at enrollment:** Age of the student at the time of enrollment.
19. **International:** Whether the student is international.
20. **Curricular units 1st sem (credited):** Number of curricular units credited in the 1st semester.
21. **Curricular units 1st sem (enrolled):** Number of curricular units enrolled in the 1st semester.
22. **Curricular units 1st sem (evaluations):** Number of curricular units evaluated in the 1st semester.
23. **Curricular units 1st sem (approved):** Number of curricular units approved in the 1st semester.
24. **Curricular units 1st sem (grade):** Grade obtained in the 1st semester.
25. **Curricular units 1st sem (without evaluations):** Number of curricular units without evaluations in the 1st semester.
26. **Curricular units 2nd sem (credited):** Number of curricular units credited in the 2nd semester.
27. **Curricular units 2nd sem (enrolled):** Number of curricular units enrolled in the 2nd semester.
28. **Curricular units 2nd sem (evaluations):** Number of curricular units evaluated in the 2nd semester.
29. **Curricular units 2nd sem (approved):** Number of curricular units approved in the 2nd semester.
30. **Curricular units 2nd sem (grade):** Grade obtained in the 2nd semester.
31. **Curricular units 2nd sem (without evaluations):** Number of curricular units without evaluations in the 2nd semester.
32. **Unemployment rate:** Unemployment rate during the student's enrollment period.
33. **Inflation rate:** Inflation rate during the student's enrollment period.
34. **GDP:** Gross Domestic Product during the student's enrollment period.

Target Variable

- **Target:** Indicates the outcome for each student:
 - "Dropout": Student dropped out.
 - "Graduate": Student successfully completed the course.

Insights and Analysis

The dataset will be analyzed to explore relationships between various features and the target variable (dropout/graduate). Visualizations will include bar plots, scatter plots, and regression plots to highlight these relationships.

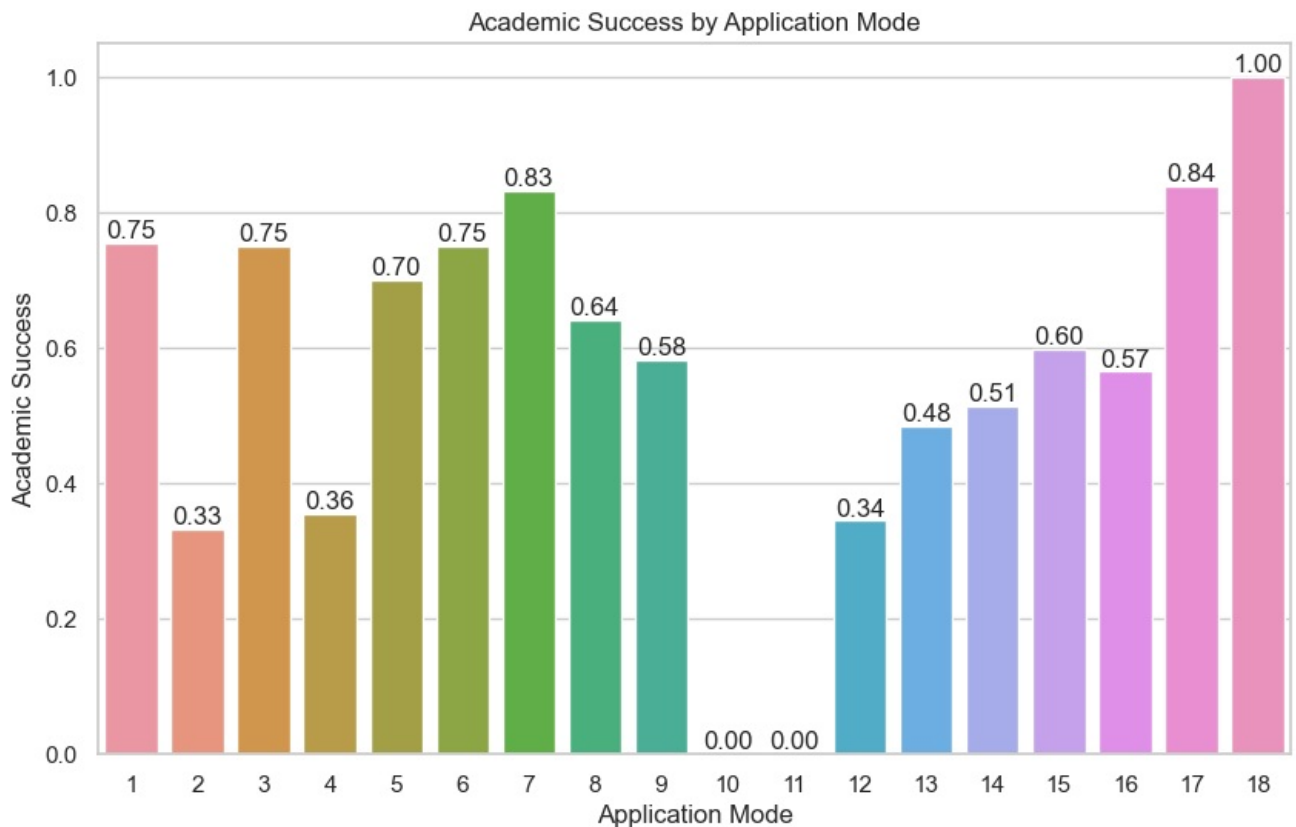
```
In [17]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from itertools import combinations
from collections import Counter
```

```
In [19]: # Load dataset
```

```
df = pd.read_csv("dataset.csv")
```

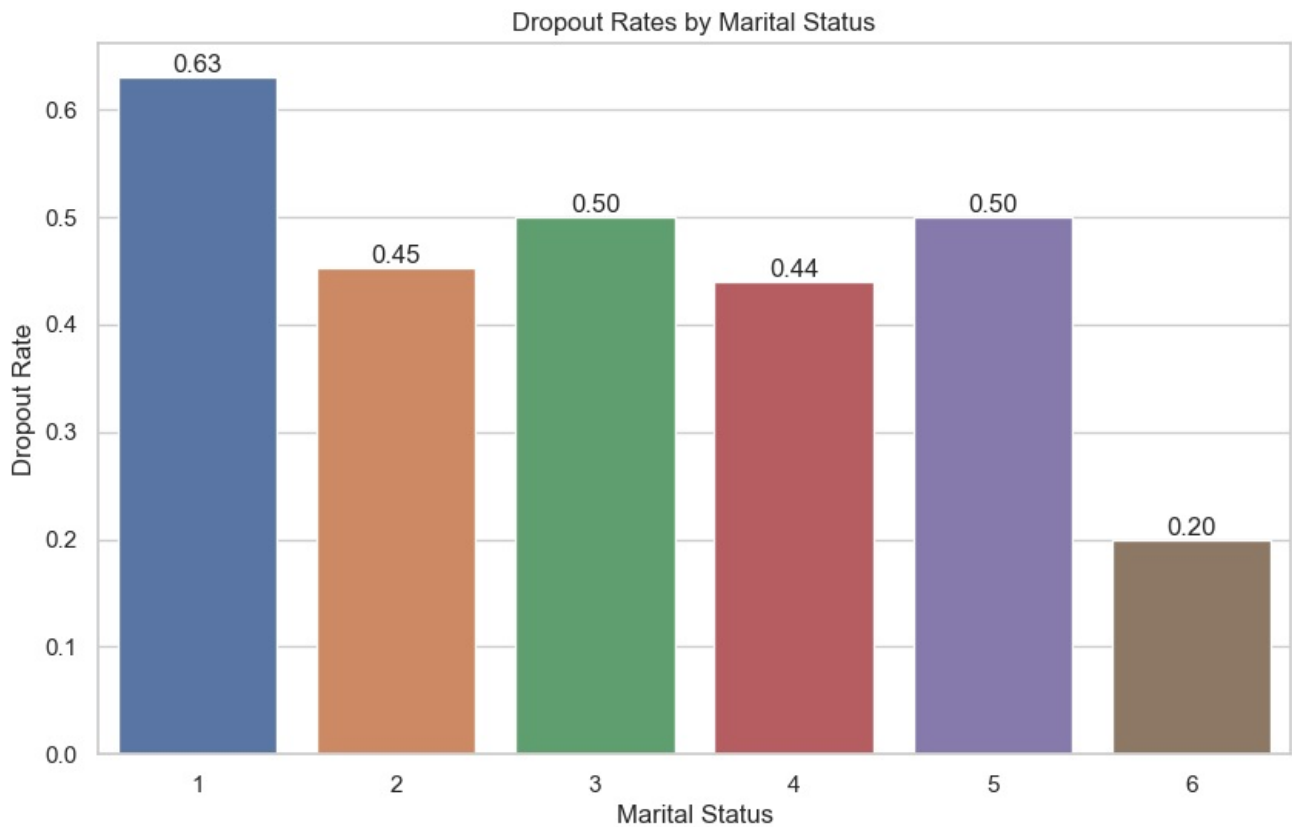
Does application mode impact academic success?

```
In [29]: # Insight 2: Application mode vs academic success
app_mode_success = df.groupby('Application mode')['Target'].mean().reset_index()
plt.figure(figsize=(10, 6))
sns.barplot(x='Application mode', y='Target', data=app_mode_success)
plt.title('Academic Success by Application Mode')
plt.xlabel('Application Mode')
plt.ylabel('Academic Success')
for i in range(len(app_mode_success)):
    plt.text(i, app_mode_success['Target'][i], f"{app_mode_success['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



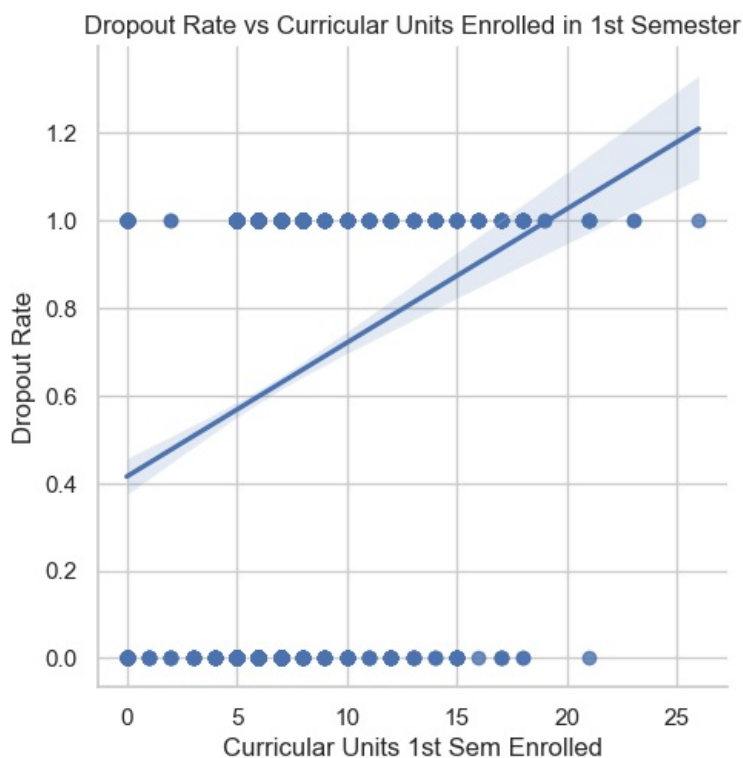
How does marital status affect dropout rates?

```
In [25]: # Insight 1: Marital status vs dropout rates
marital_dropout = df.groupby('Marital status')['Target'].mean().reset_index()
plt.figure(figsize=(10, 6))
sns.barplot(x='Marital status', y='Target', data=marital_dropout)
plt.title('Dropout Rates by Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Dropout Rate')
for i in range(len(marital_dropout)):
    plt.text(i, marital_dropout['Target'][i], f"{marital_dropout['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



What is the relationship between the number of curricular units enrolled in the 1st semester and dropout rates?

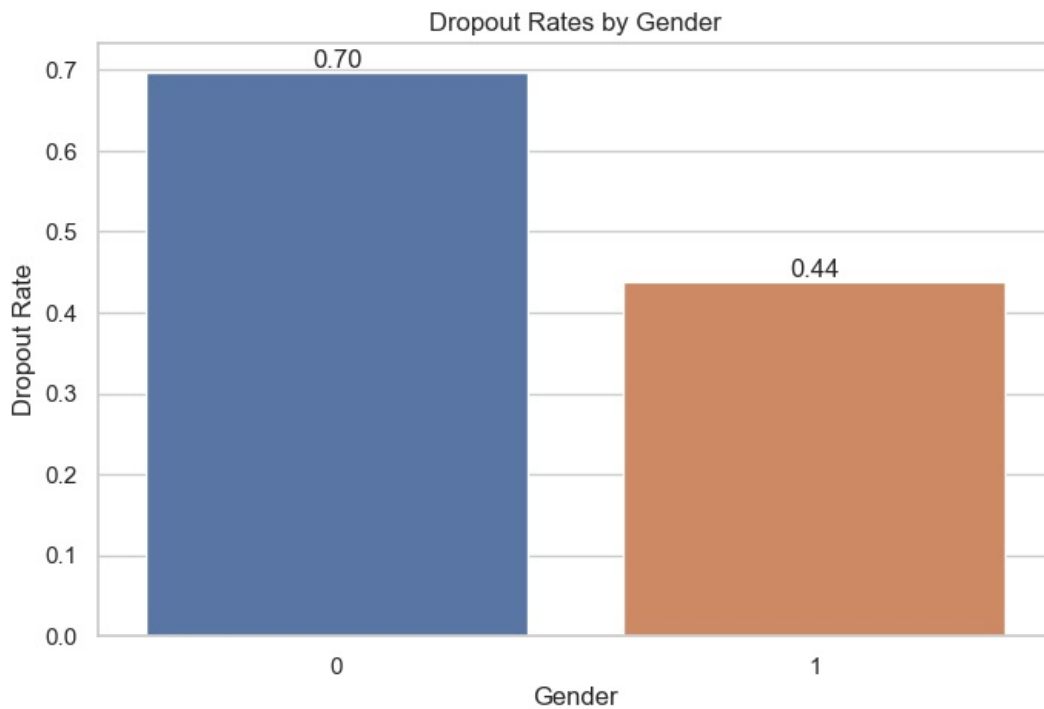
```
In [34]: # Insight 3: Curricular units 1st sem enrolled vs dropout rates
sns.lmplot(x='Curricular units 1st sem (enrolled)', y='Target', data=df)
plt.title('Dropout Rate vs Curricular Units Enrolled in 1st Semester')
plt.xlabel('Curricular Units 1st Sem Enrolled')
plt.ylabel('Dropout Rate')
plt.show()
```



Does gender play a role in dropout rates?

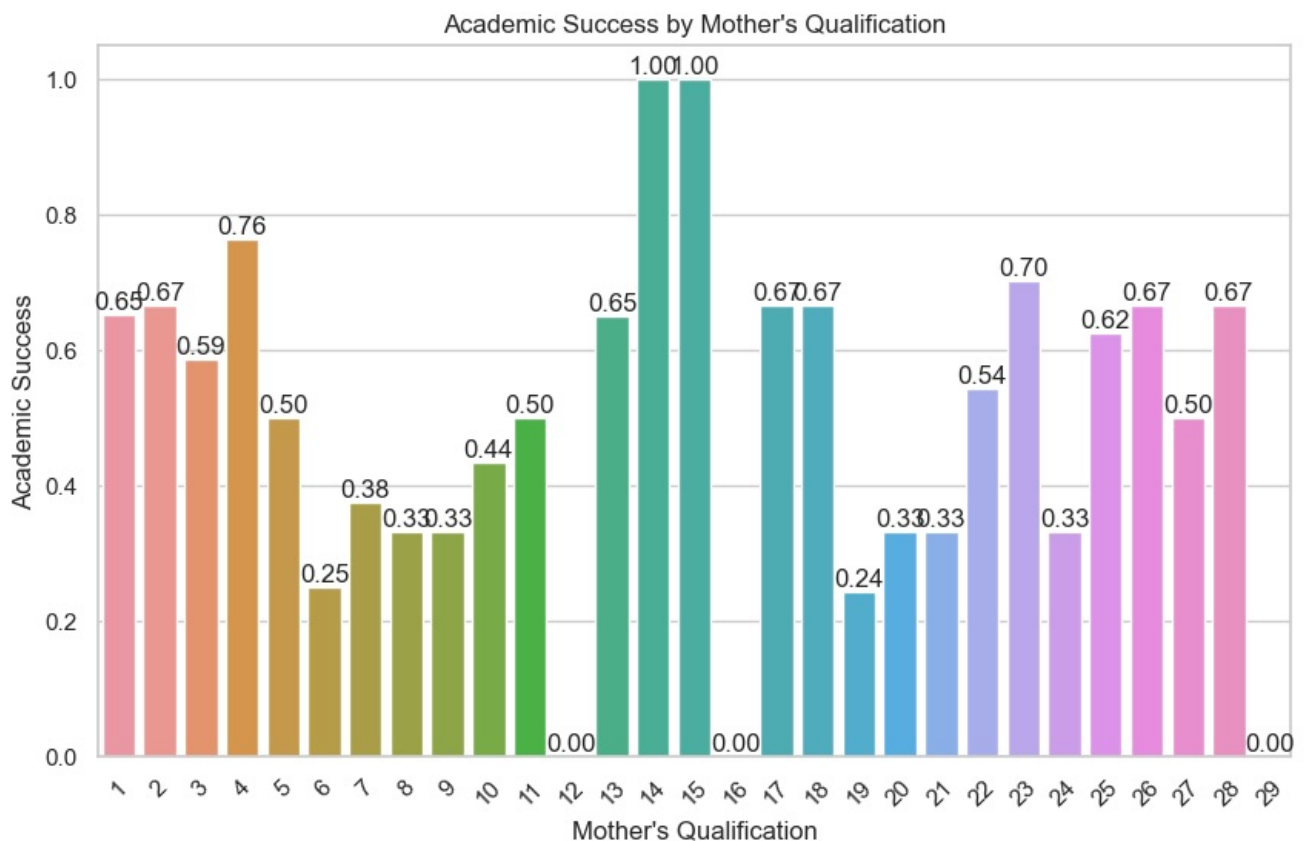
```
In [37]: # Insight 4: Gender vs dropout rates
gender_dropout = df.groupby('Gender')['Target'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(x='Gender', y='Target', data=gender_dropout)
plt.title('Dropout Rates by Gender')
```

```
plt.xlabel('Gender')
plt.ylabel('Dropout Rate')
for i in range(len(gender_dropout)):
    plt.text(i, gender_dropout['Target'][i], f"{gender_dropout['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



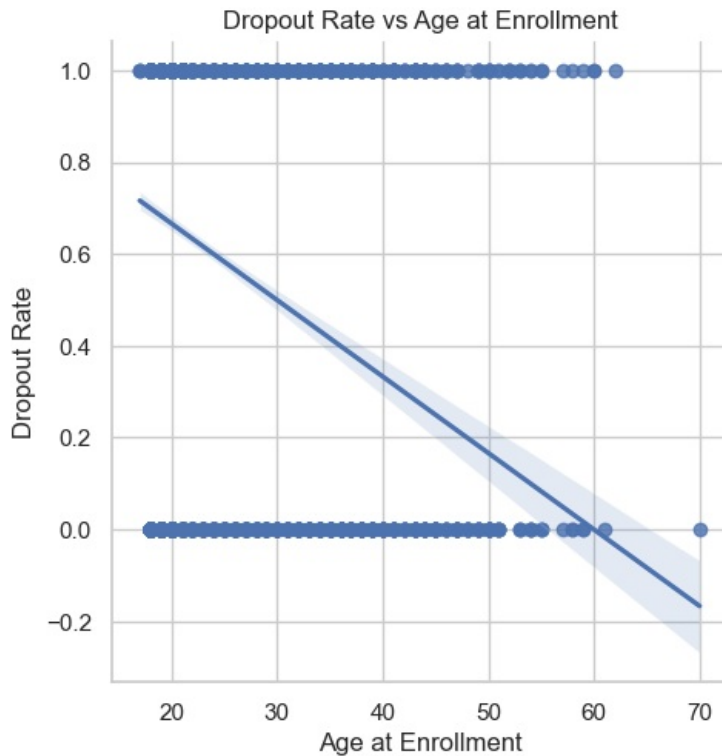
How does the educational qualification of the mother affect academic success?

```
In [40]: # Insight 5: Mother's qualification vs academic success
mother_qual_success = df.groupby('Mother\'s qualification')['Target'].mean().reset_index()
plt.figure(figsize=(10, 6))
sns.barplot(x='Mother\'s qualification', y='Target', data=mother_qual_success)
plt.title('Academic Success by Mother's Qualification')
plt.xlabel('Mother\'s Qualification')
plt.ylabel('Academic Success')
plt.xticks(rotation=45)
for i in range(len(mother_qual_success)):
    plt.text(i, mother_qual_success['Target'][i], f"{mother_qual_success['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



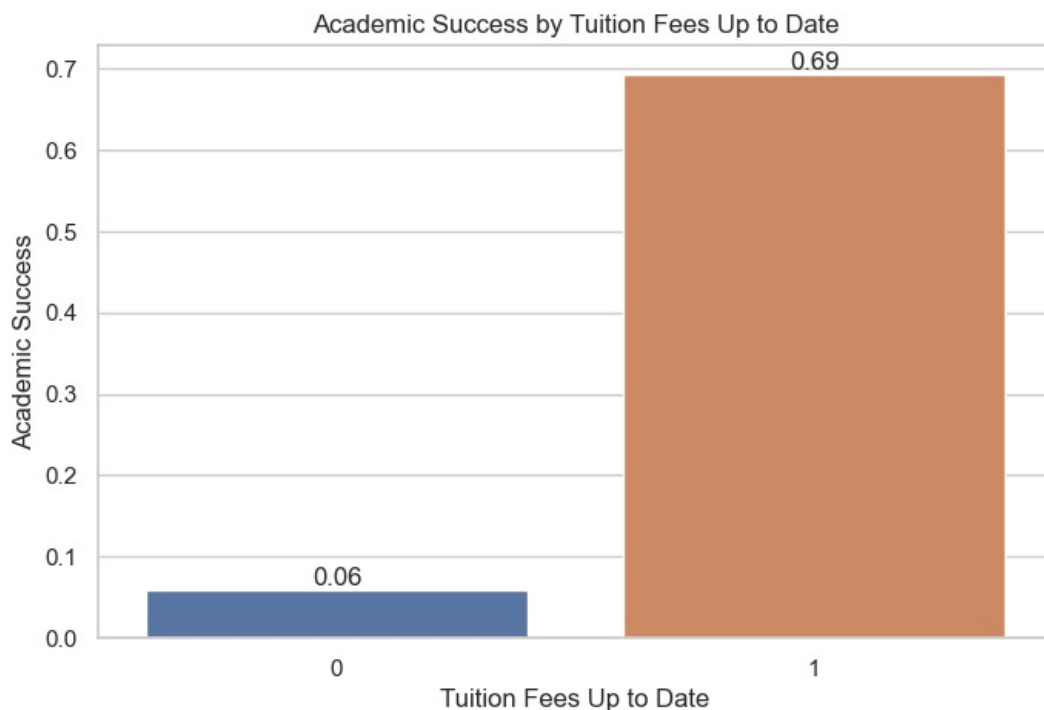
Insight 6: Age at enrollment vs dropout ratesow()

```
In [43]: # Insight 6: Age at enrollment vs dropout rates
sns.lmplot(x='Age at enrollment', y='Target', data=df)
plt.title('Dropout Rate vs Age at Enrollment')
plt.xlabel('Age at Enrollment')
plt.ylabel('Dropout Rate')
plt.show()
```



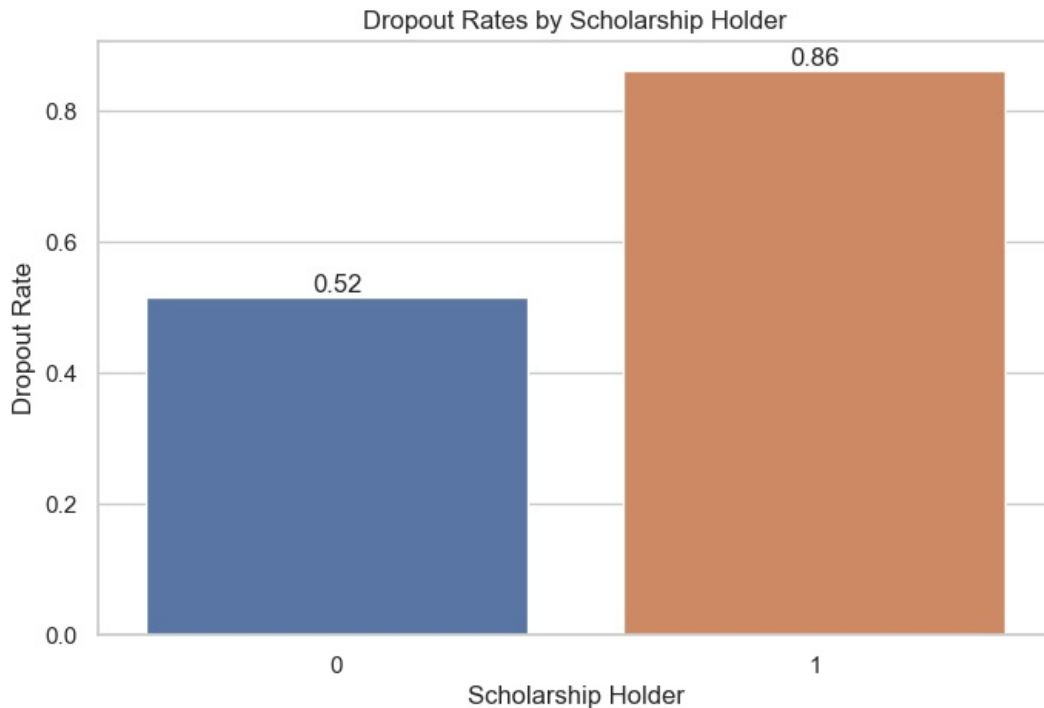
Is there a correlation between tuition fees being up to date and academic success?

```
In [46]: # Insight 7: Tuition fees up to date vs academic success
tuition_success = df.groupby('Tuition fees up to date')['Target'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(x='Tuition fees up to date', y='Target', data=tuition_success)
plt.title('Academic Success by Tuition Fees Up to Date')
plt.xlabel('Tuition Fees Up to Date')
plt.ylabel('Academic Success')
for i in range(len(tuition_success)):
    plt.text(i, tuition_success['Target'][i], f"{tuition_success['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



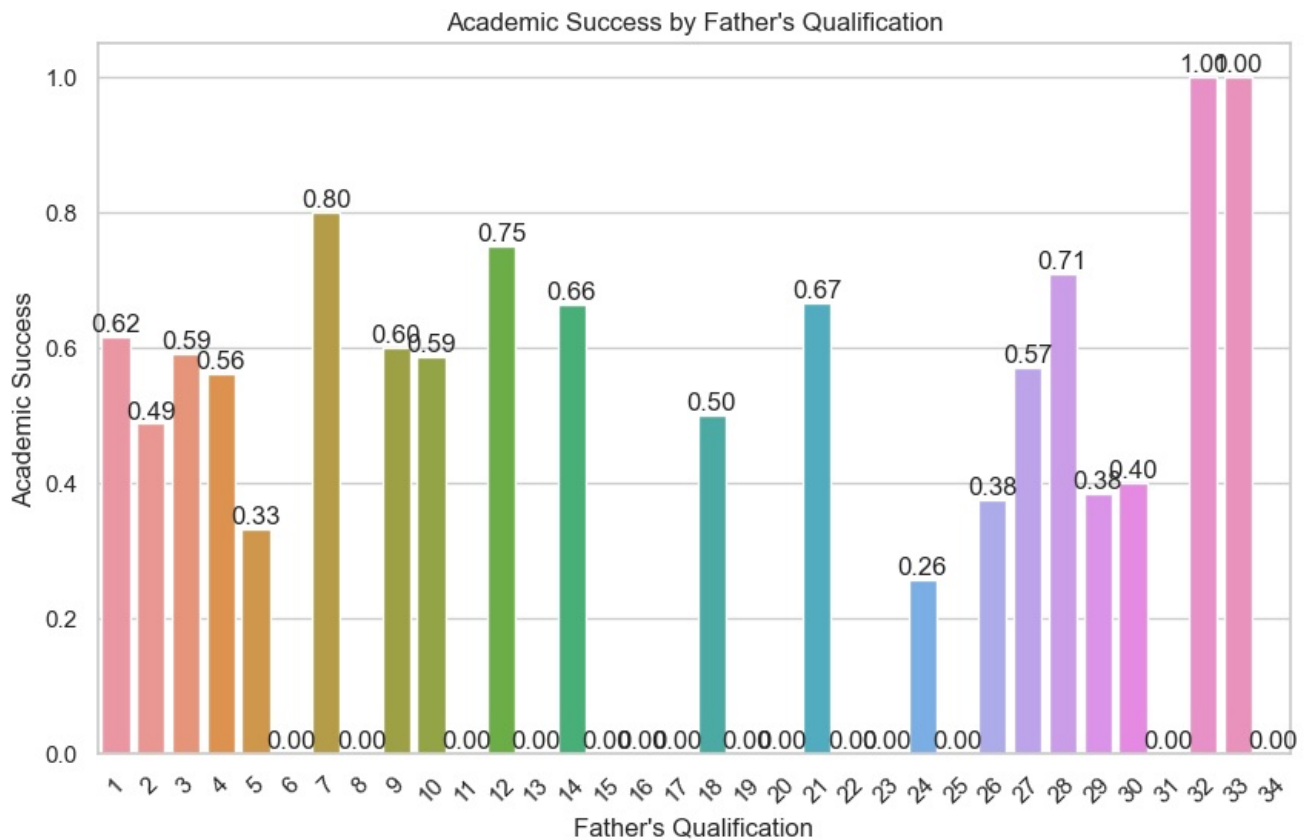
How does being a scholarship holder affect dropout rates?

```
In [48]: # Insight 8: Scholarship holder vs dropout rates
scholarship_dropout = df.groupby('Scholarship holder')['Target'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(x='Scholarship holder', y='Target', data=scholarship_dropout)
plt.title('Dropout Rates by Scholarship Holder')
plt.xlabel('Scholarship Holder')
plt.ylabel('Dropout Rate')
for i in range(len(scholarship_dropout)):
    plt.text(i, scholarship_dropout['Target'][i], f"{scholarship_dropout['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



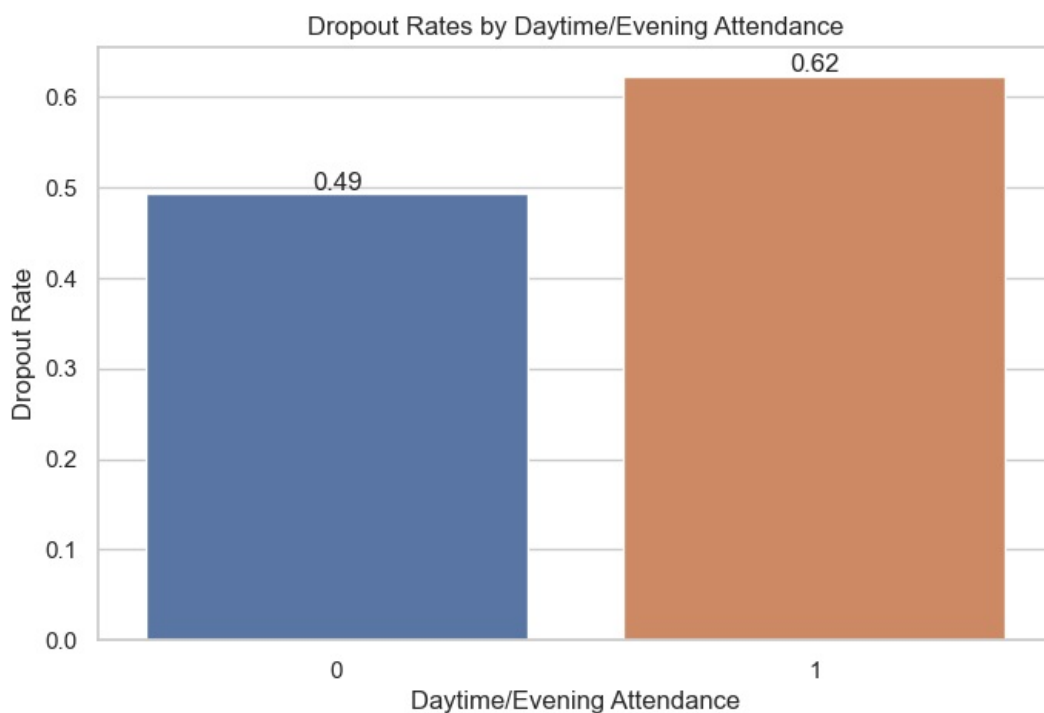
Does the father's qualification have any impact on academic success?

```
In [52]: # Insight 9: Father's qualification vs academic success
father_qual_success = df.groupby('Father\'s qualification')['Target'].mean().reset_index()
plt.figure(figsize=(10, 6))
sns.barplot(x='Father\'s qualification', y='Target', data=father_qual_success)
plt.title('Academic Success by Father\'s Qualification')
plt.xlabel('Father\'s Qualification')
plt.ylabel('Academic Success')
plt.xticks(rotation=45)
for i in range(len(father_qual_success)):
    plt.text(i, father_qual_success['Target'][i], f"{father_qual_success['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



What is the effect of daytime/evening attendance on dropout rates?

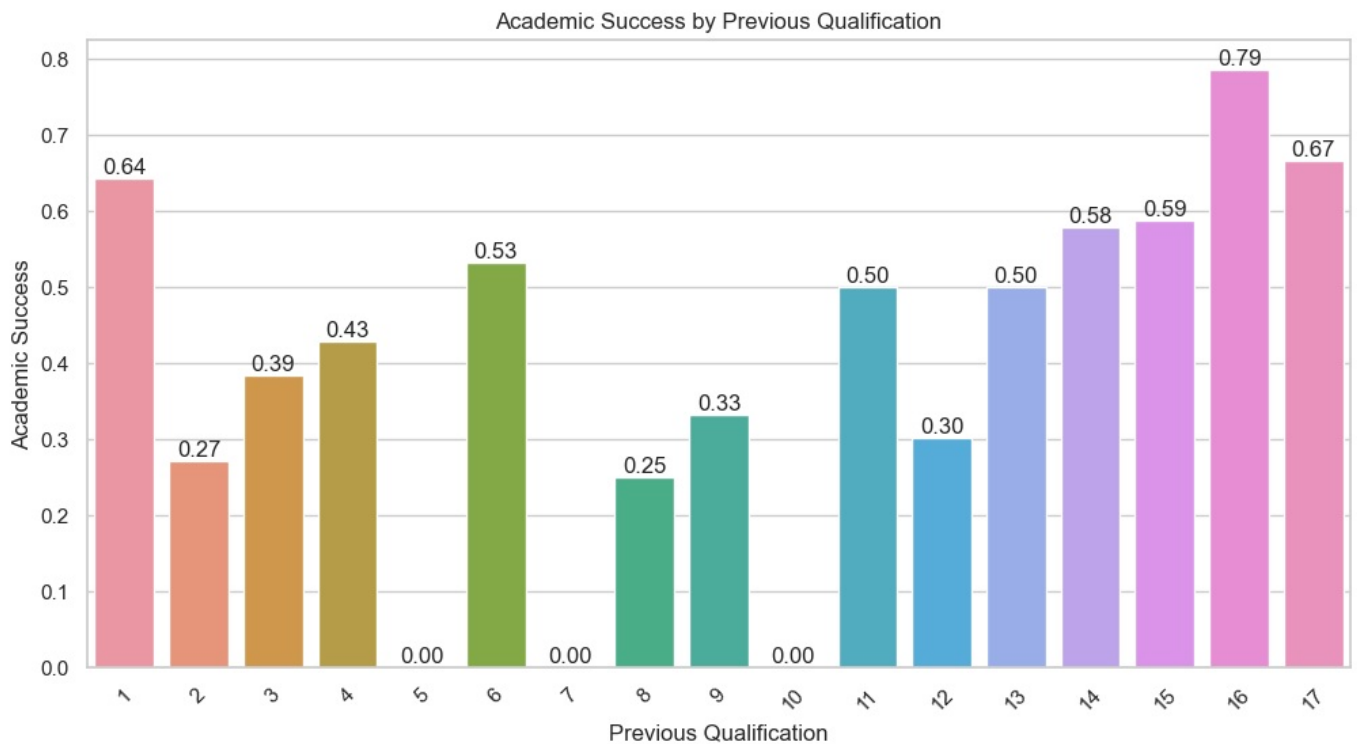
```
In [54]: # Insight 10: Daytime/evening attendance vs dropout rates
attendance_dropout = df.groupby('Daytime/evening attendance')['Target'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(x='Daytime/evening attendance', y='Target', data=attendance_dropout)
plt.title('Dropout Rates by Daytime/Evening Attendance')
plt.xlabel('Daytime/Evening Attendance')
plt.ylabel('Dropout Rate')
for i in range(len(attendance_dropout)):
    plt.text(i, attendance_dropout['Target'][i], f"{attendance_dropout['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



How do previous qualifications influence academic success?

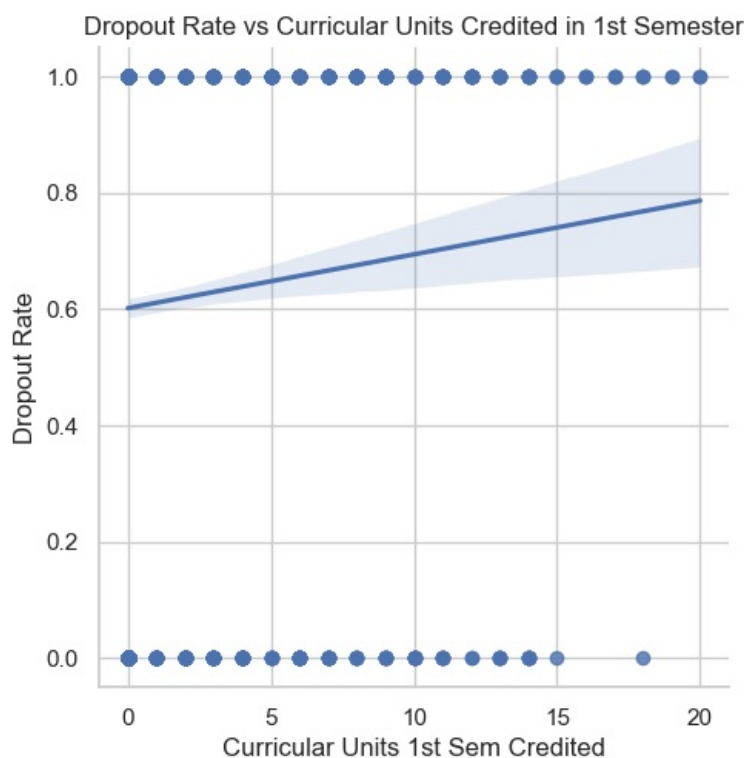
```
In [57]: # Insight 11: Previous qualification vs academic success
prev_qual_success = df.groupby('Previous qualification')['Target'].mean().reset_index()
plt.figure(figsize=(12, 6))
sns.barplot(x='Previous qualification', y='Target', data=prev_qual_success)
```

```
plt.title('Academic Success by Previous Qualification')
plt.xlabel('Previous Qualification')
plt.ylabel('Academic Success')
plt.xticks(rotation=45)
for i in range(len(prev_qual_success)):
    plt.text(i, prev_qual_success['Target'][i], f"{prev_qual_success['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()
```



Is there a relationship between the number of curricular units credited in the 1st semester and dropout rates?

```
In [61]: # Insight 12: Curricular units 1st sem credited vs dropout rates
sns.lmplot(x='Curricular units 1st sem (credited)', y='Target', data=df)
plt.title('Dropout Rate vs Curricular Units Credited in 1st Semester')
plt.xlabel('Curricular Units 1st Sem Credited')
plt.ylabel('Dropout Rate')
plt.show()
```



How does being displaced affect dropout rates?

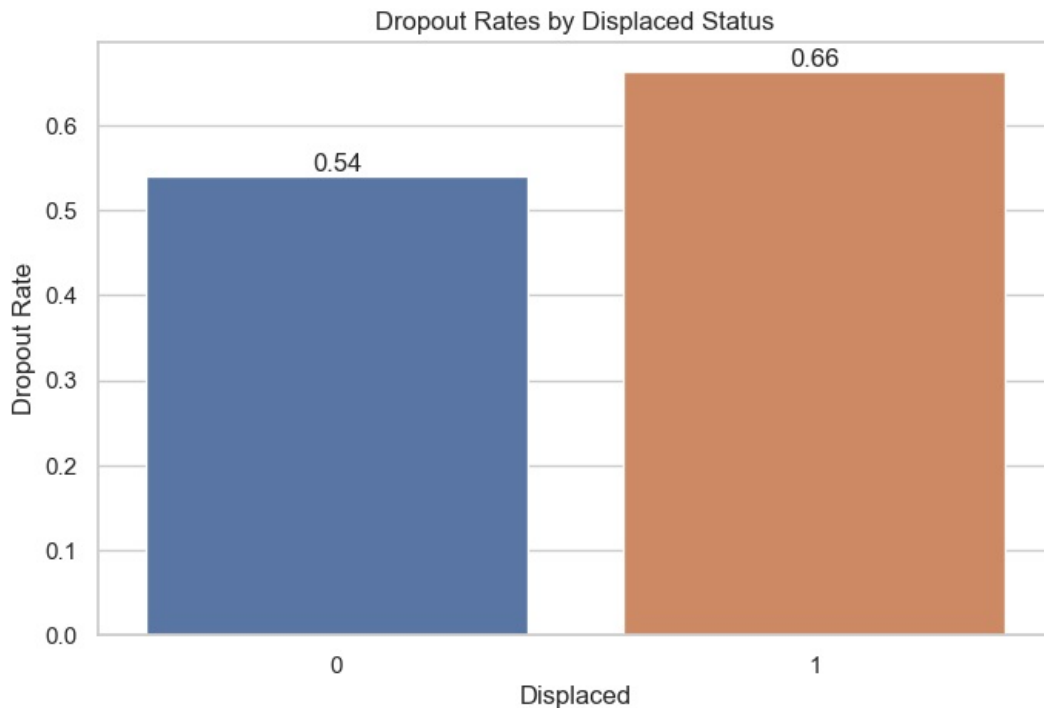
```
In [64]: # Insight 13: Displaced vs dropout rates
```



```

displaced_dropout = df.groupby('Displaced')['Target'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(x='Displaced', y='Target', data=displaced_dropout)
plt.title('Dropout Rates by Displaced Status')
plt.xlabel('Displaced')
plt.ylabel('Dropout Rate')
for i in range(len(displaced_dropout)):
    plt.text(i, displaced_dropout['Target'][i], f"{displaced_dropout['Target'][i]:.2f}", ha='center', va='bottom')
plt.show()

```

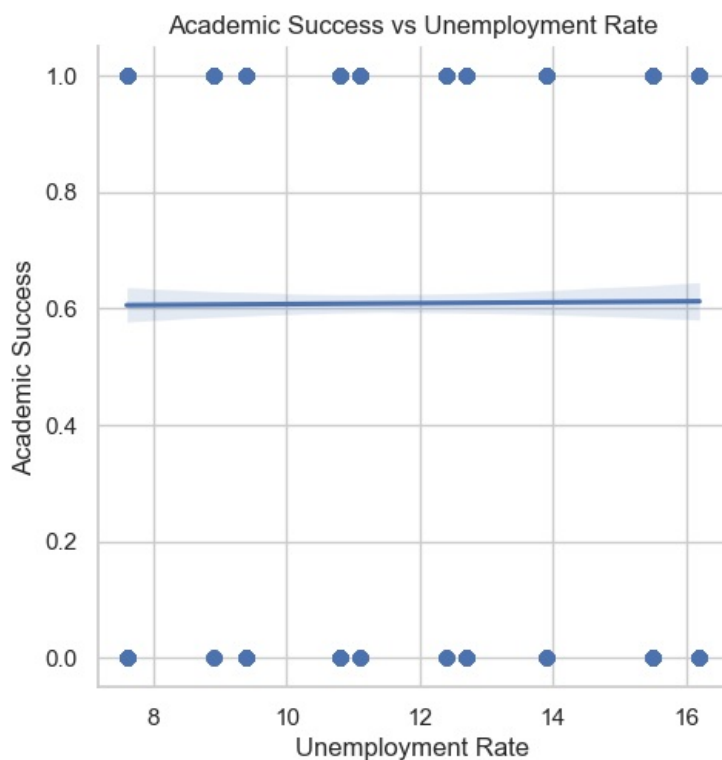


Does the unemployment rate correlate with academic success?

```

In [66]: # Insight 14: Unemployment rate vs academic success
sns.lmplot(x='Unemployment rate', y='Target', data=df)
plt.title('Academic Success vs Unemployment Rate')
plt.xlabel('Unemployment Rate')
plt.ylabel('Academic Success')
plt.show()

```



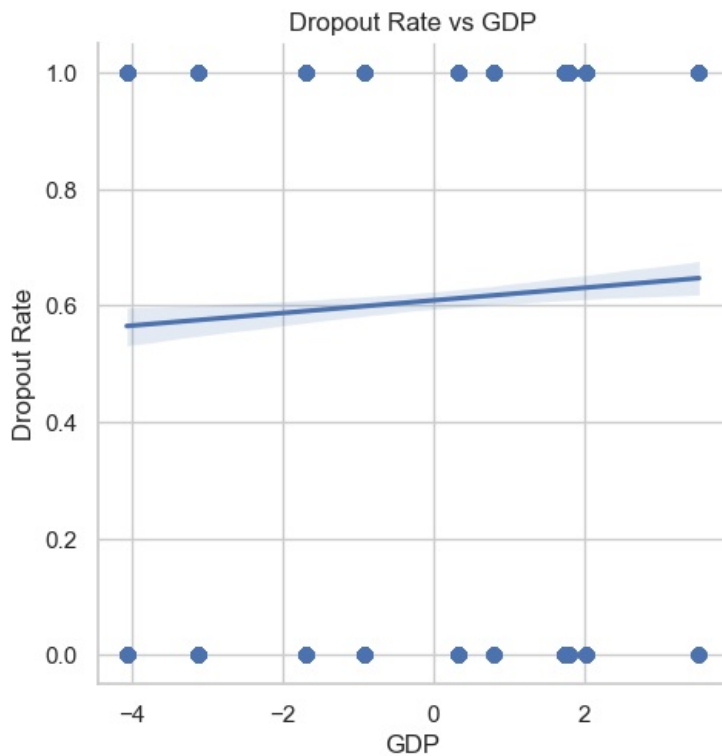
Is there a correlation between GDP and dropout rates?

```

In [70]: # Insight 15: GDP vs dropout rates

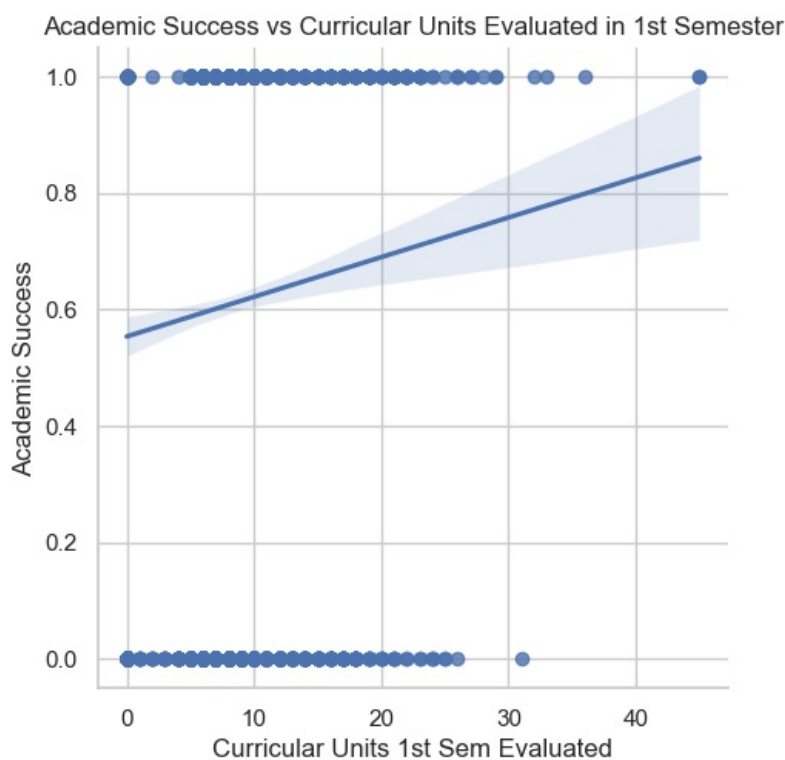
```

```
sns.lmplot(x='GDP', y='Target', data=df)
plt.title('Dropout Rate vs GDP')
plt.xlabel('GDP')
plt.ylabel('Dropout Rate')
plt.show()
```



Is there a relationship between the number of curricular units evaluated and academic success in the 1st semester?

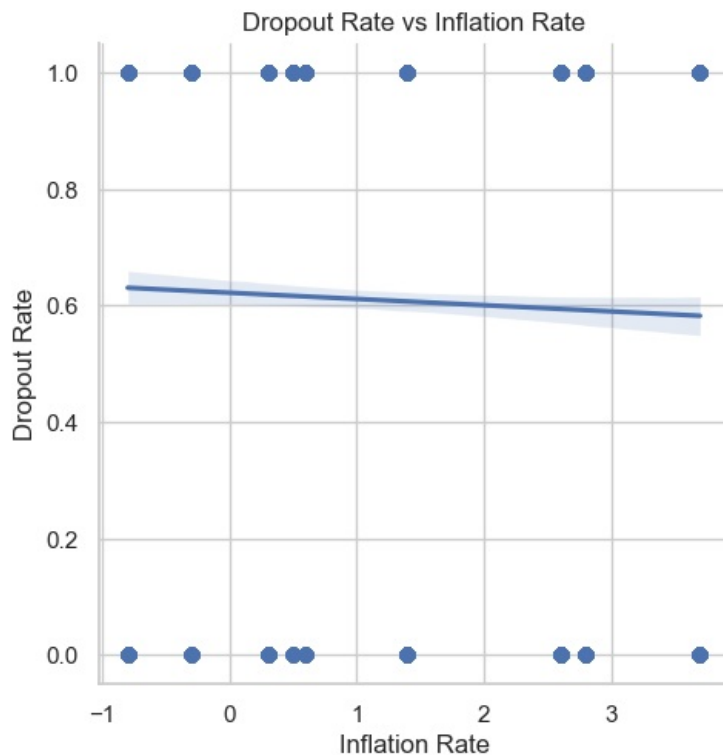
```
In [76]: # Insight 17: Curricular units 1st sem evaluated vs academic success
sns.lmplot(x='Curricular units 1st sem (evaluations)', y='Target', data=df)
plt.title('Academic Success vs Curricular Units Evaluated in 1st Semester')
plt.xlabel('Curricular Units 1st Sem Evaluated')
plt.ylabel('Academic Success')
plt.show()
```



Does the inflation rate during enrollment impact dropout rates?

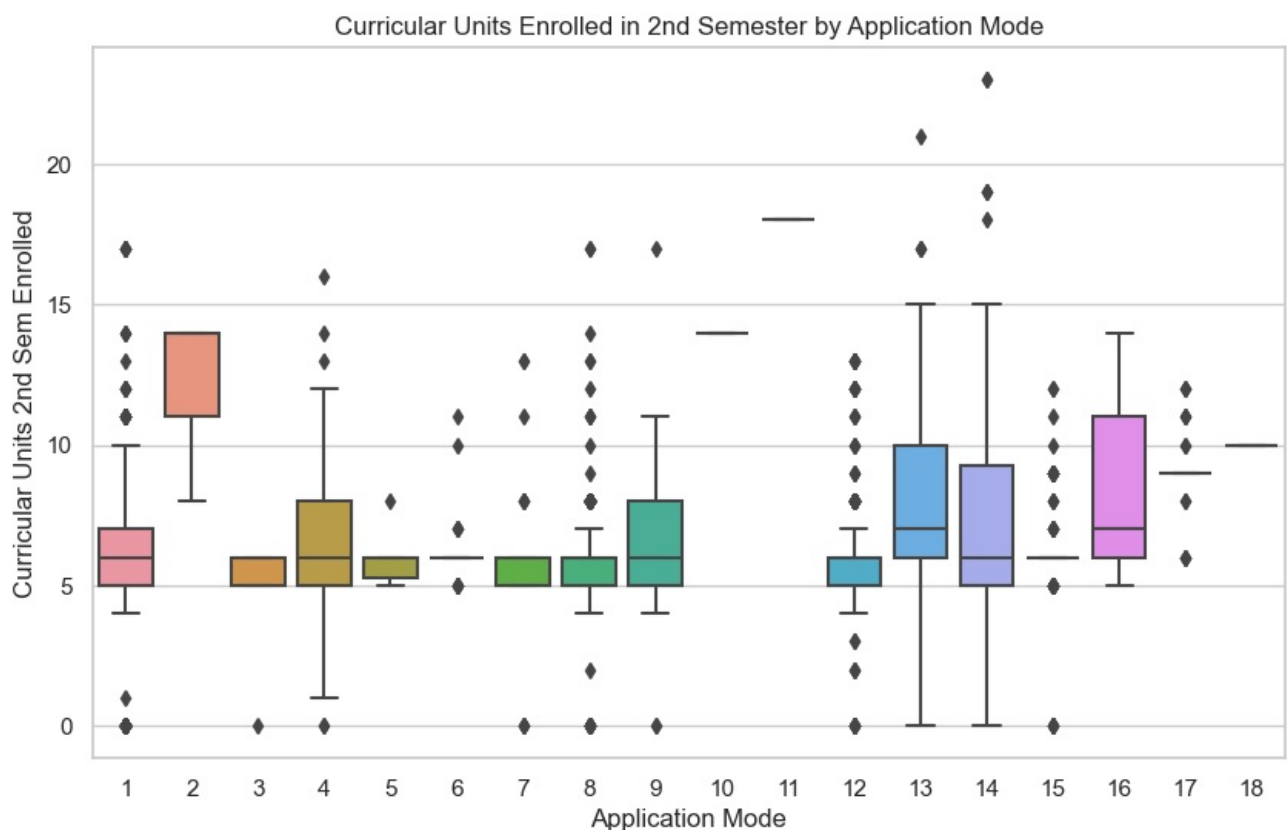
```
In [80]: # Insight 18: Inflation rate vs dropout rates
sns.lmplot(x='Inflation rate', y='Target', data=df)
```

```
plt.title('Dropout Rate vs Inflation Rate')
plt.xlabel('Inflation Rate')
plt.ylabel('Dropout Rate')
plt.show()
```



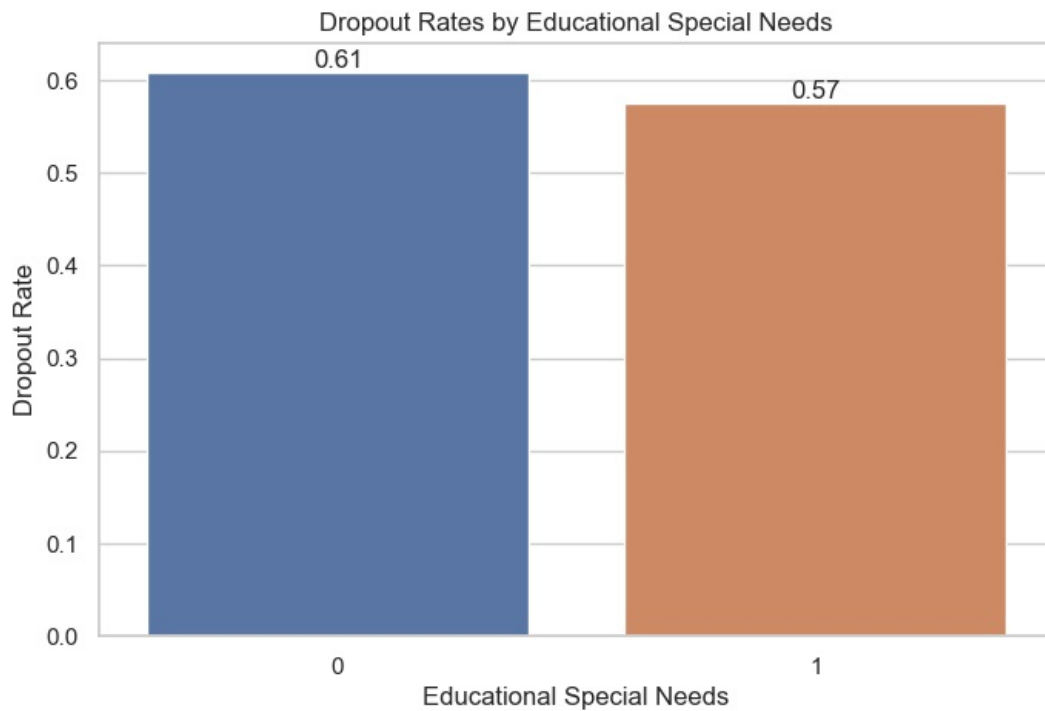
How does the type of application mode affect the number of curricular units enrolled in the 2nd semester?

```
In [85]: # Insight 19: Application mode vs curricular units 2nd sem enrolled
plt.figure(figsize=(10, 6))
sns.boxplot(x='Application mode', y='Curricular units 2nd sem (enrolled)', data=df)
plt.title('Curricular Units Enrolled in 2nd Semester by Application Mode')
plt.xlabel('Application Mode')
plt.ylabel('Curricular Units 2nd Sem Enrolled')
plt.show()
```



Is there a difference in dropout rates between students with and without educational special needs?

```
In [88]: # Insight 20: Educational special needs vs dropout rates
special_needs_dropout = df.groupby('Educational special needs')['Target'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(x='Educational special needs', y='Target', data=special_needs_dropout)
plt.title('Dropout Rates by Educational Special Needs')
plt.xlabel('Educational Special Needs')
plt.ylabel('Dropout Rate')
for i in range(len(special_needs_dropout)):
    plt.text(i, special_needs_dropout['Target'][i], f"{special_needs_dropout['Target'][i]:.2f}", ha='center', va='top')
plt.show()
```



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js