eda

October 24, 2024

1 EDA Analysis

1.1 Background

The dataset sourced from the UCI Machine Learning Repo contains 4424 student entries with 36 features and one categorical target varaible. The goal is to predict student academic success (dropout or graduate) from the 36 input features.

[2]:

```
!pip install ucimlrepo  --quiet
from visualization_utils import *
from file_utils import *
import matplotlib.pyplot as plt
from ucimlrepo import fetch_ucirepo
```

[3]:

```
# fetch dataset
predict_students_dropout_and_academic_success = fetch_ucirepo(id=697)

# data (as pandas dataframes)
X = predict_students_dropout_and_academic_success.data.features
y = predict_students_dropout_and_academic_success.data.targets

# metadata
metadata = predict_students_dropout_and_academic_success.metadata

# variable information
variable_info = predict_students_dropout_and_academic_success.variables
df = X
df['dropout'] = y
print(f"{df.shape[0]} entries with {df.shape[1]} features")
```

```
4424 entries with 37 features
```

[4]:

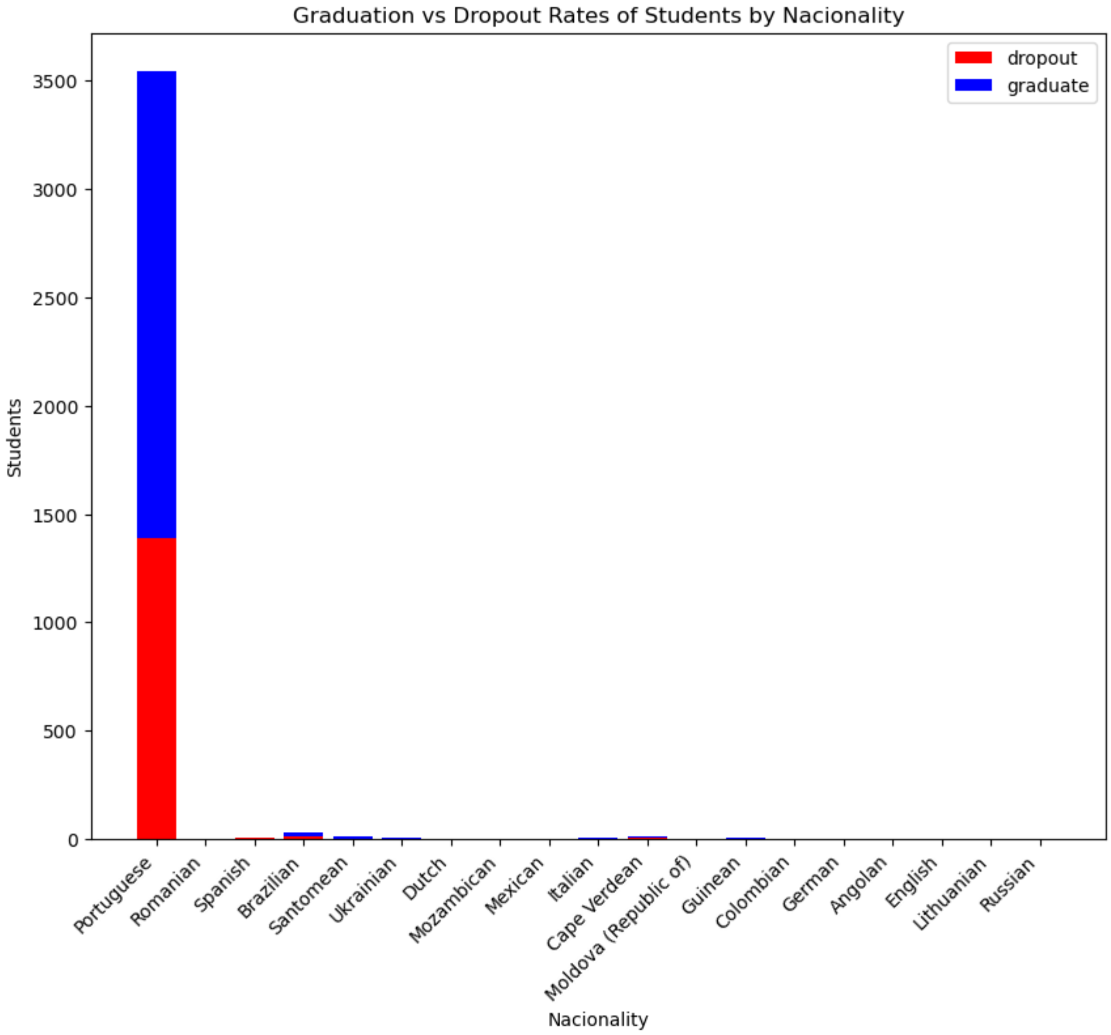
```
predict_students_dropout_and_academic_success.variables['name']
ds_vars = predict_students_dropout_and_academic_success.variables['name']
ds_desc = predict_students_dropout_and_academic_success.variables['description']
# need to create forward mappings for each one of the variables
quantitative_vars = {'Application mode',
'Application order',
'Previous qualification (grade)',
'Admission grade',
'Age at enrollment',
"Curricular units 1st sem (credited)",
"Curricular units 1st sem (enrolled)",
"Curricular units 1st sem (evaluations)",
"Curricular units 1st sem (approved)",
"Curricular units 1st sem (grade)",
"Curricular units 1st sem (without evaluations)",
"Curricular units 2nd sem (credited)",
"Curricular units 2nd sem (enrolled)",
"Curricular units 2nd sem (evaluations)",
"Curricular units 2nd sem (approved)",
"Curricular units 2nd sem (grade)",
"Curricular units 2nd sem (without evaluations)",
"Unemployment rate",
"Inflation rate",
"GDP",
"Target",
}

variable_map = generate_variable_map(quantitative_vars, ds_vars, ds_desc)
```

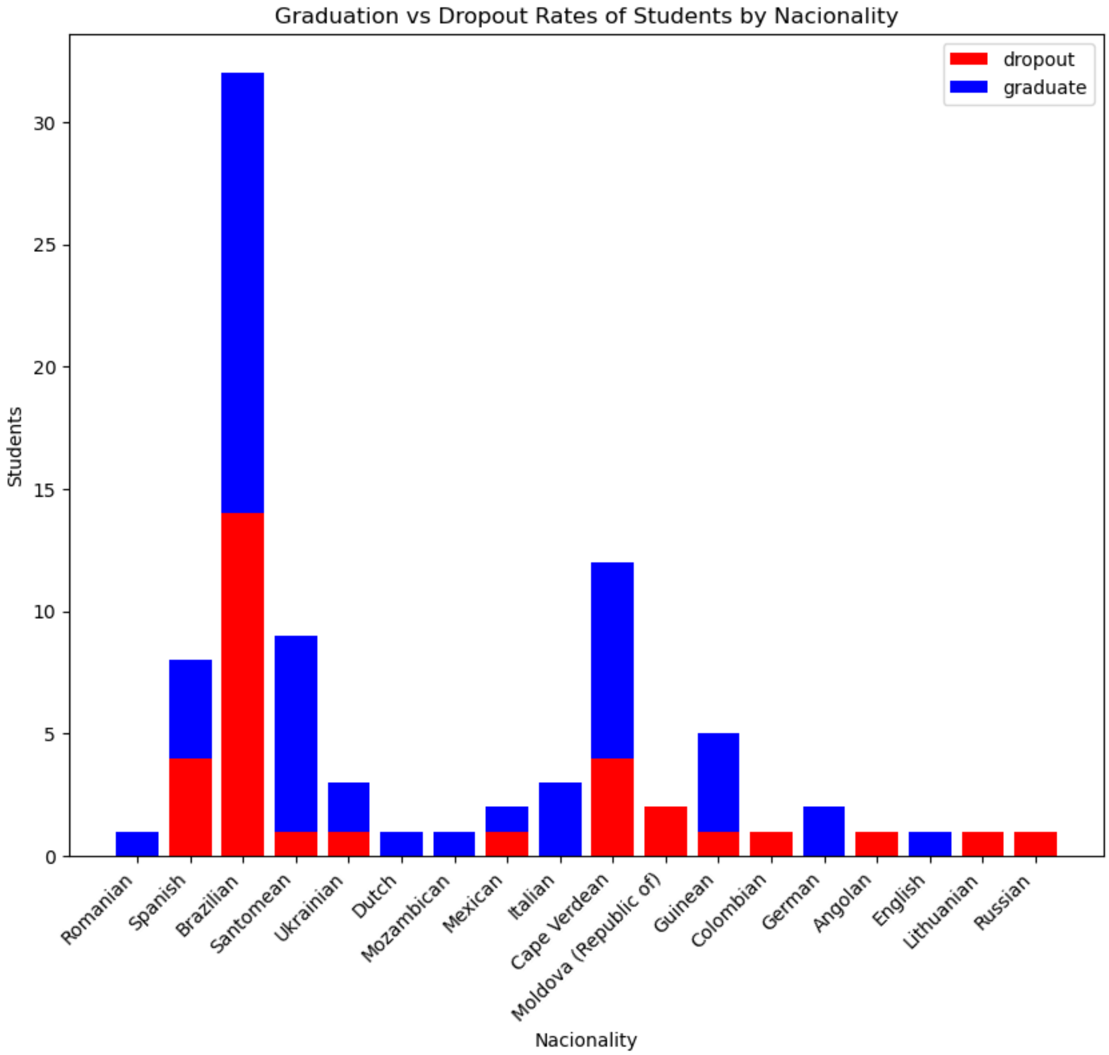
2 Analysis of Categorical Variables Affecting Student Performance

To gain a better understanding of variables responsibel for controlling the

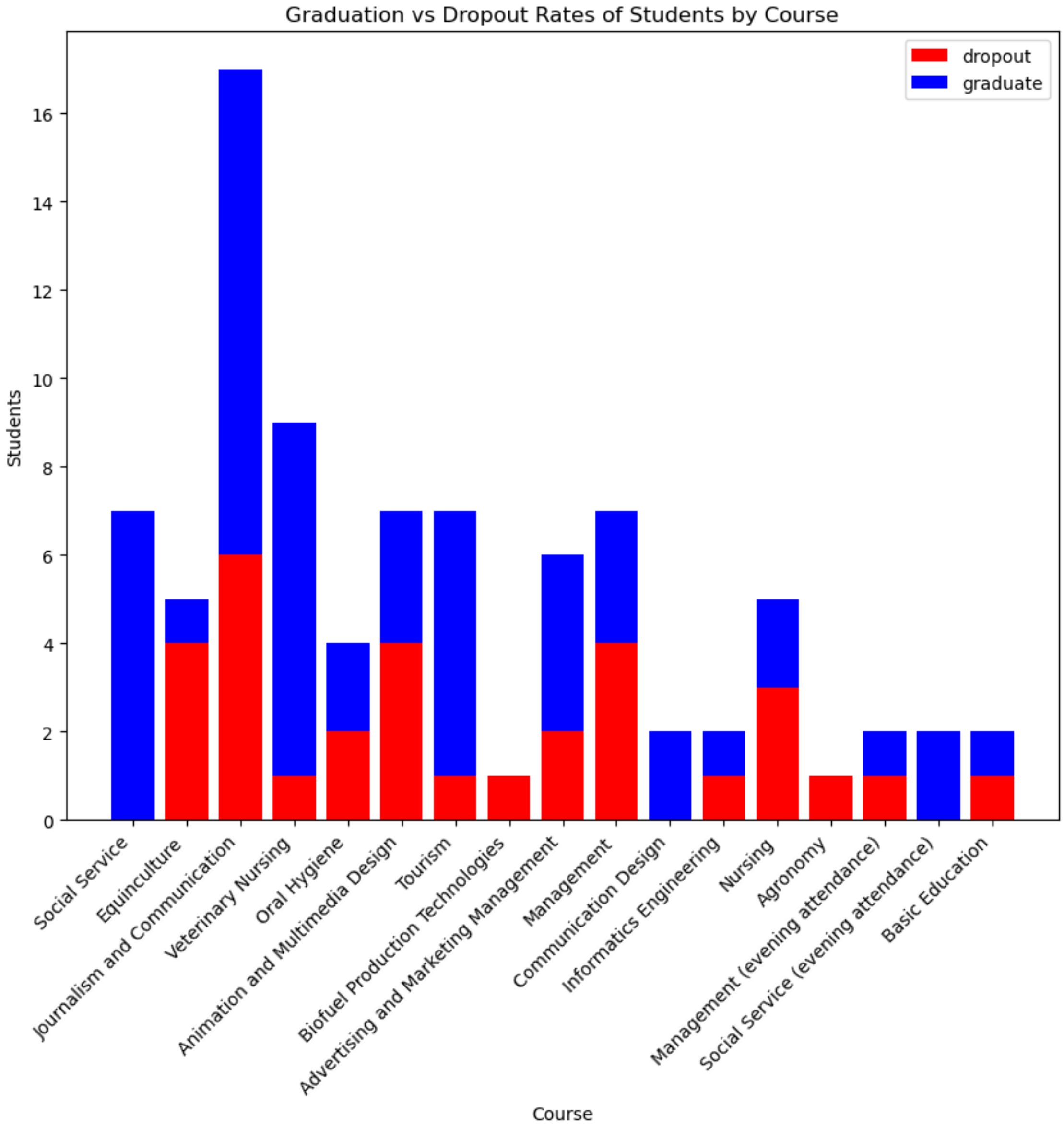
[5]:

```
finished_df = df[df['dropout'].isin(['Dropout', 'Graduate'])]
generate_general_stacked_bar_graph(finished_df, variable_map, 'Nacionality')
finished_df = finished_df.drop(finished_df.loc[finished_df['Nacionality']==1].
  ↪index)
generate_general_stacked_bar_graph(finished_df, variable_map, 'Nacionality')
```

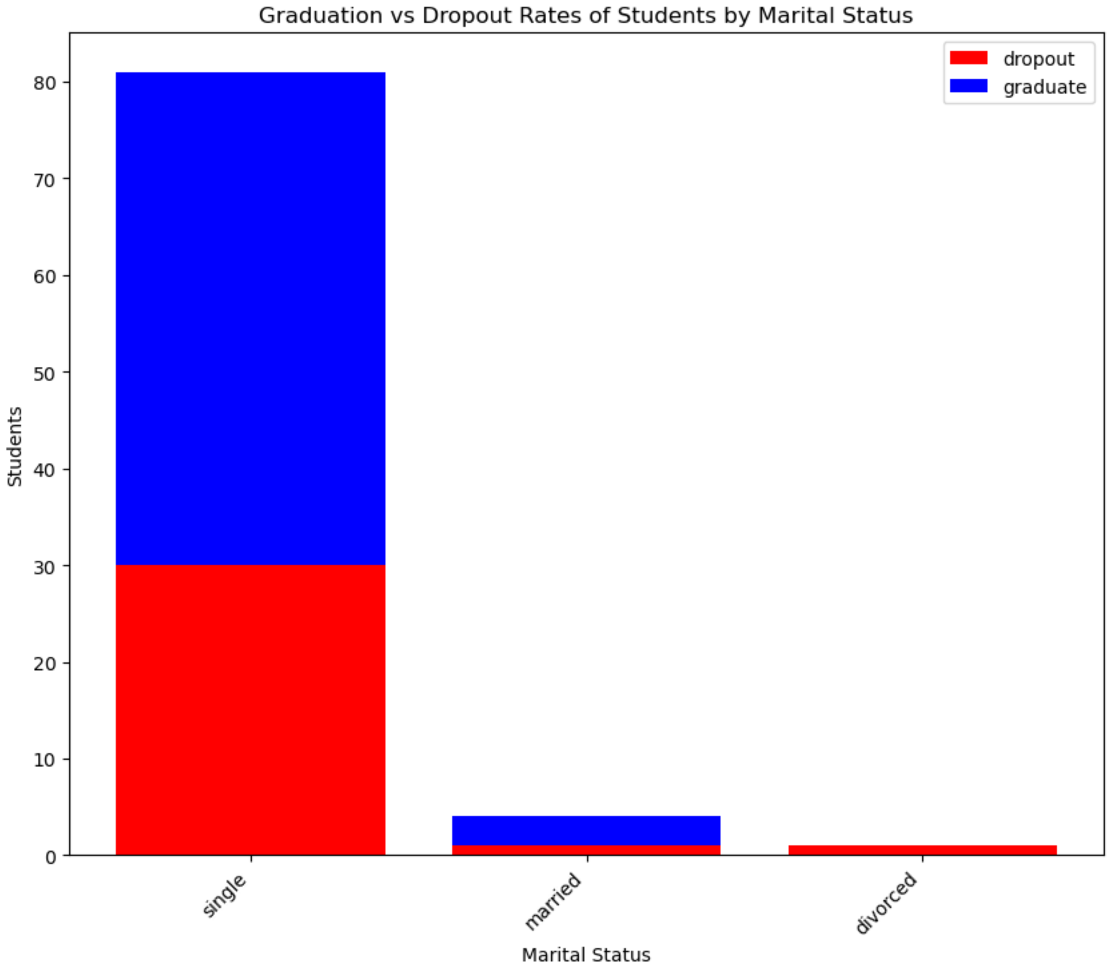

Graduation vs Dropout Rates of Students by Nacionality
dropout
graduate
Students
3500
3000
2500
2000
1500
1000
500
0
Portuguese
Romanian
Spanish
Brazilian
Santomean
Ukrainian
Dutch
Mozambican
Mexican
Italian
Cape Verdean
Moldova (Republic of)
Guinean
Colombian
German
Angolan
English
Lithuanian
Russian
Nacionality



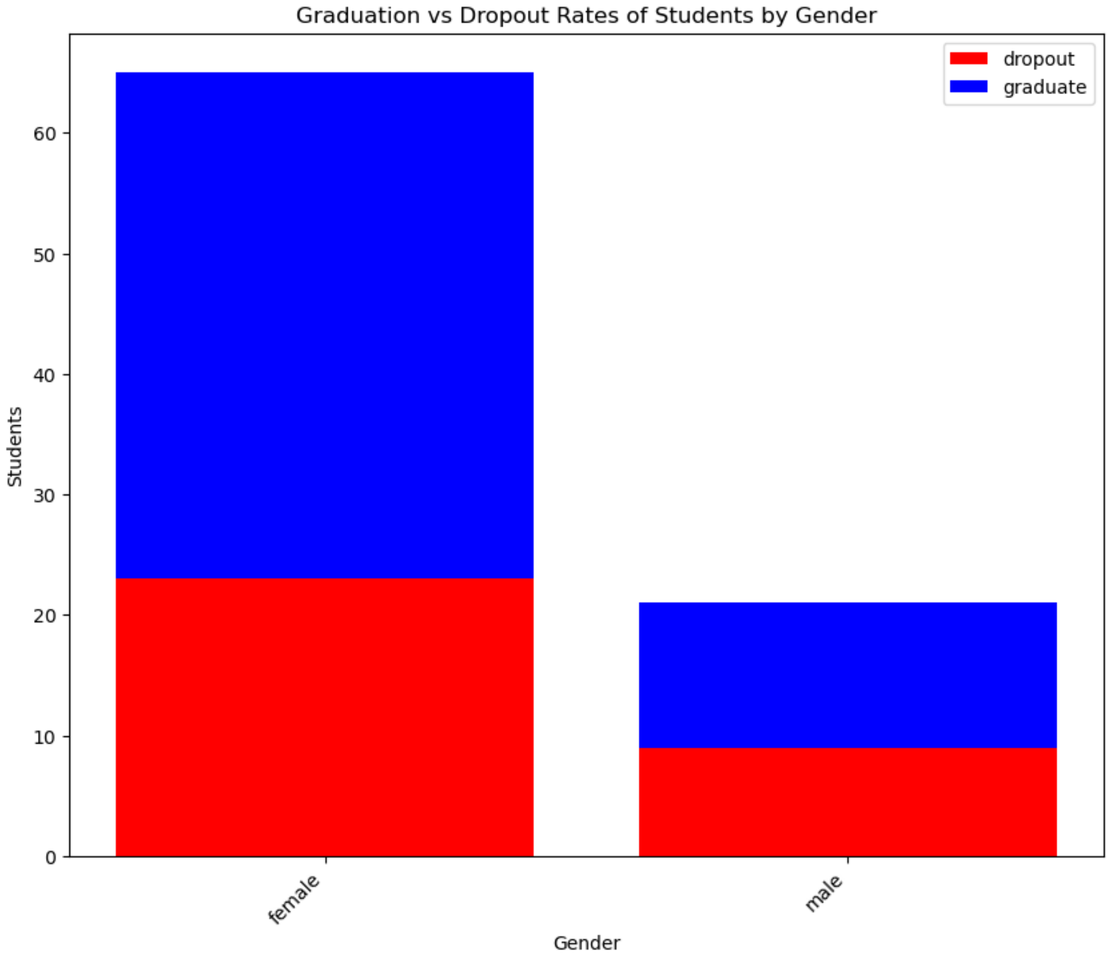
```
[6]: generate_general_stacked_bar_graph(finished_df, variable_map, 'Course')
```



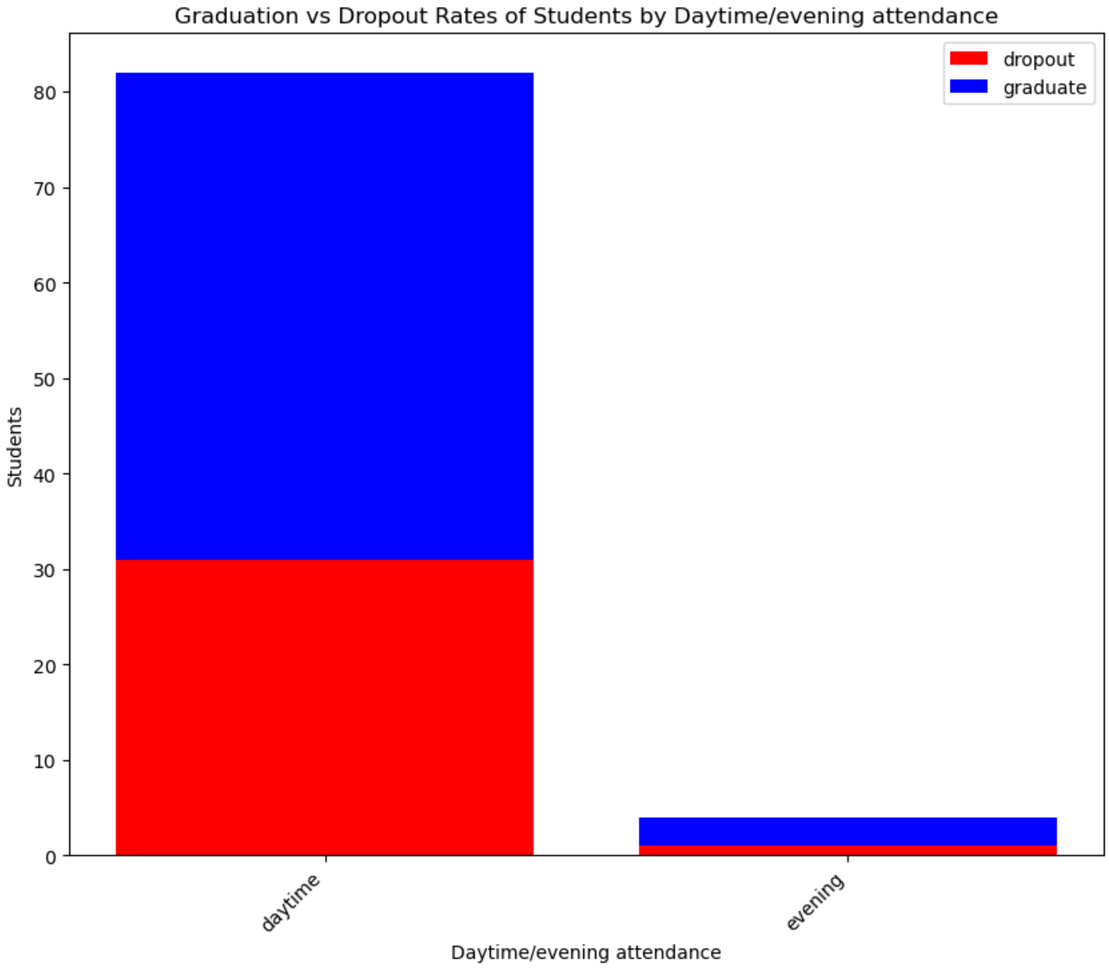
[7]:
```
generate_general_stacked_bar_graph(finished_df, variable_map, 'Marital Status')
```



[8]:
```
generate_general_stacked_bar_graph(finished_df, variable_map, 'Gender')
```

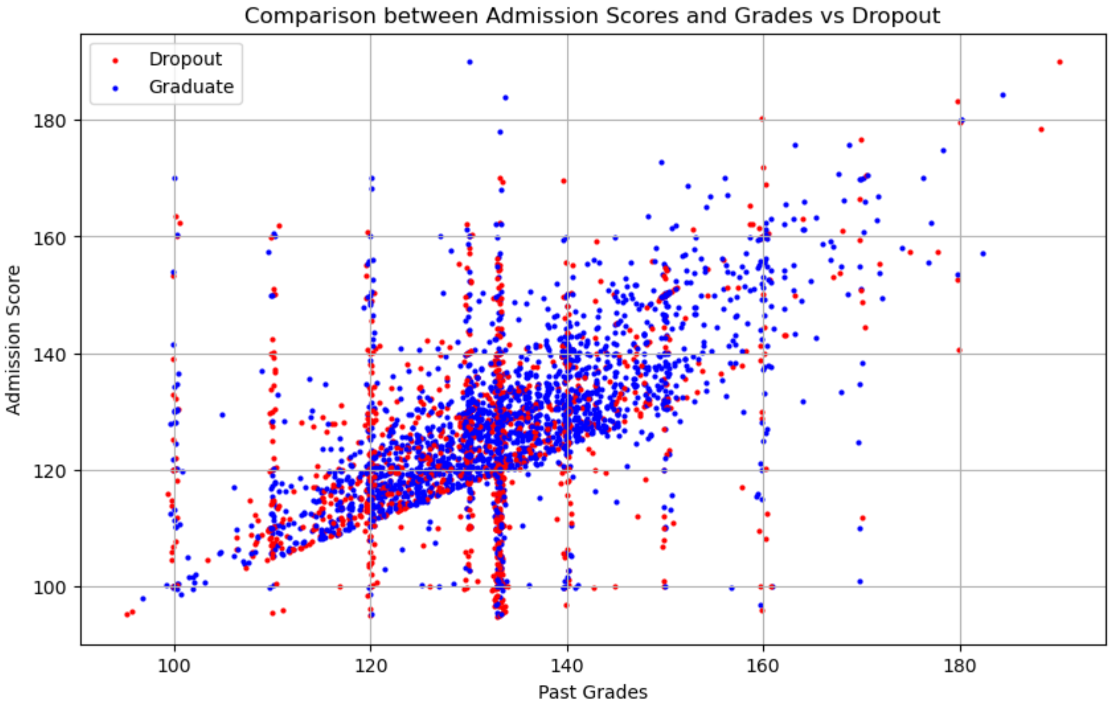


```
[9]: generate_general_stacked_bar_graph(finished_df, variable_map, 'Daytime/evening attendance')
```



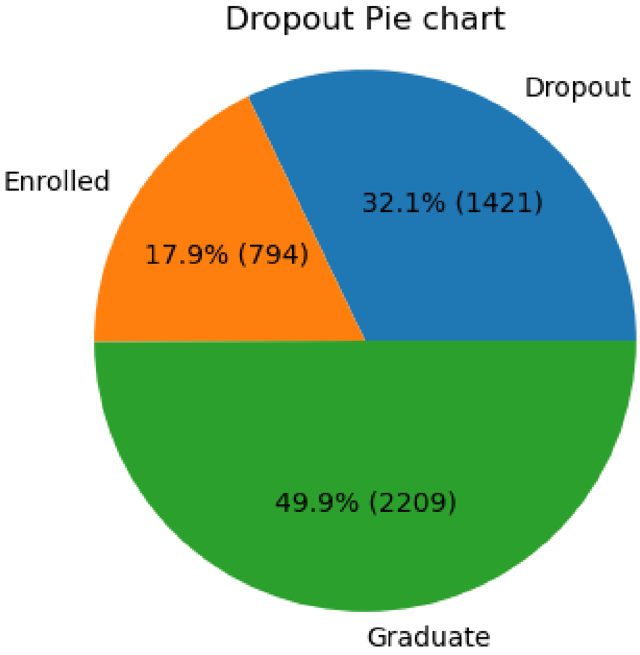
[10]:

```
PREV_QUAL = 'Previous qualification (grade)'
ADMISSION_GRADE = 'Admission grade'
tmp = df[[PREV_QUAL, ADMISSION_GRADE, 'dropout']]
finished_df = tmp[tmp['dropout'].isin(['Dropout', 'Graduate'])]

jitter_plot(finished_df)
```



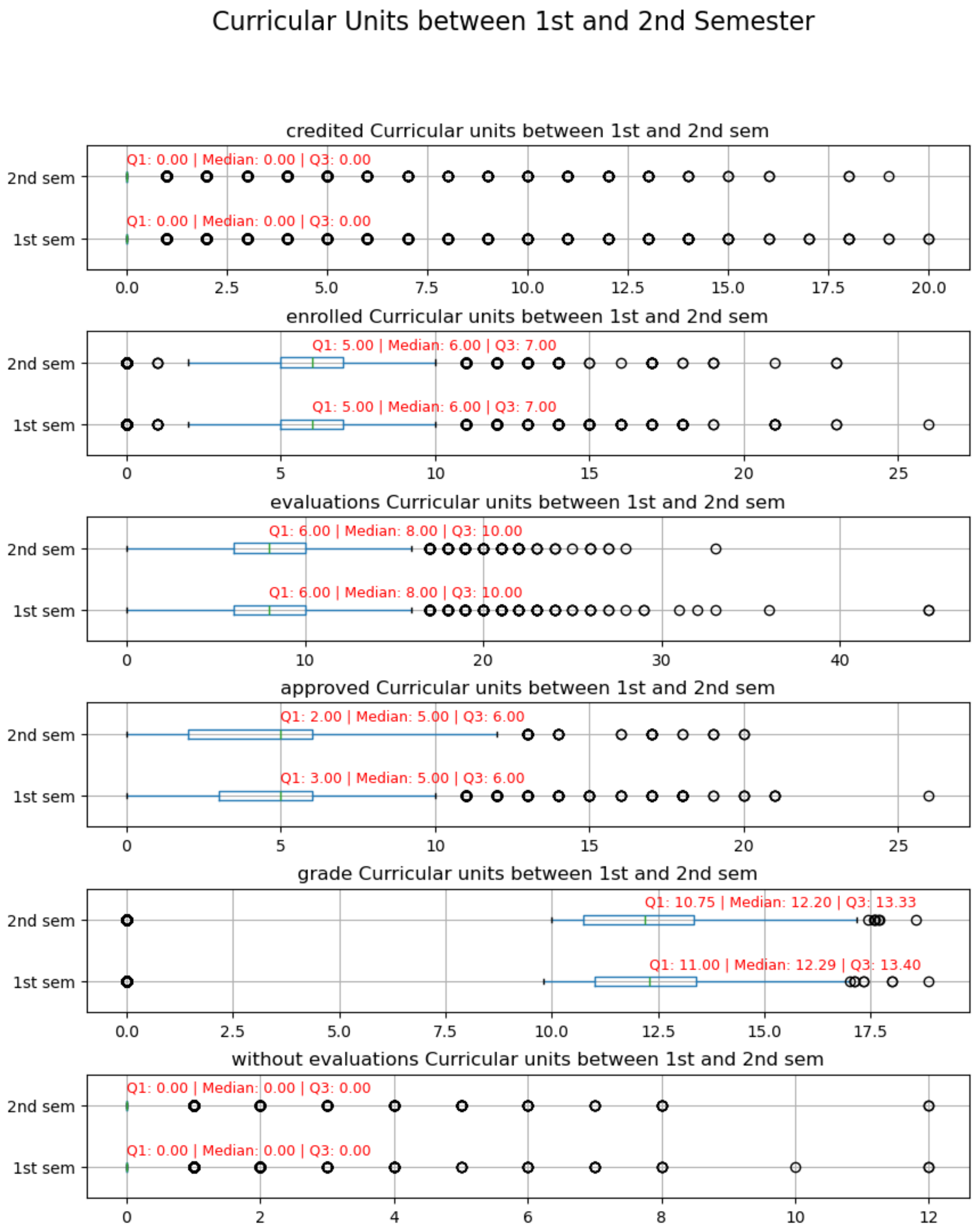
[11]:

```
pie_chart(df['dropout'])
```



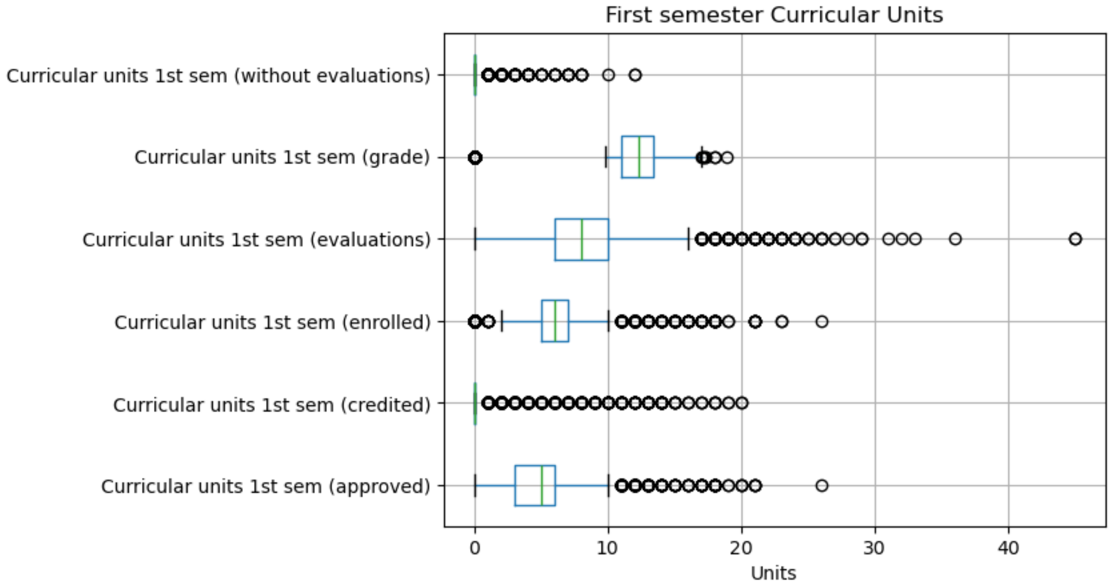
[12]:

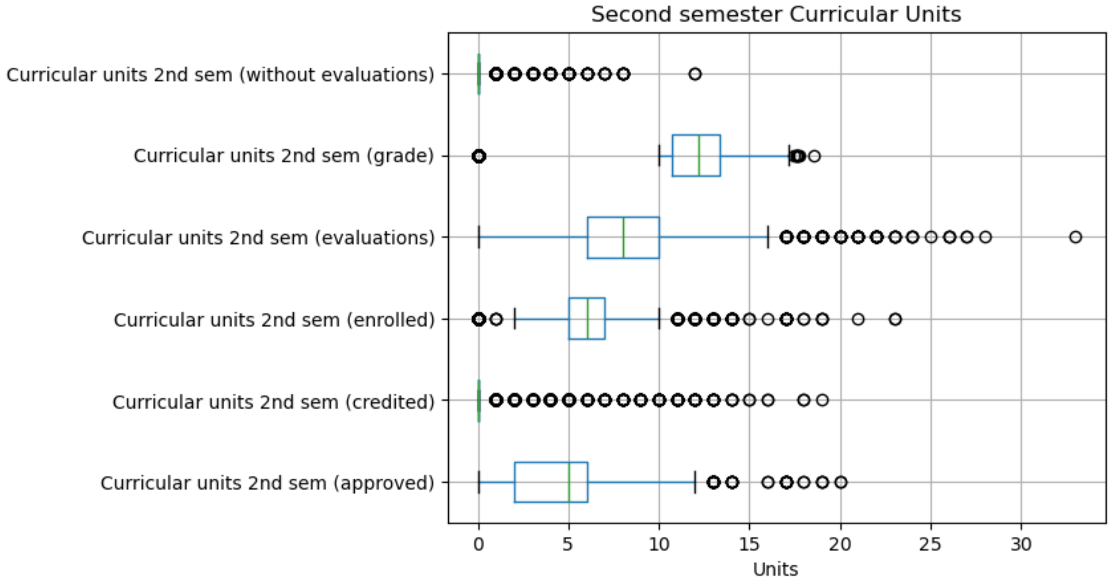
```python
curriculum_units = [
    "Curricular units 1st sem (credited)",
    "Curricular units 2nd sem (credited)",
    "Curricular units 1st sem (enrolled)",
    "Curricular units 2nd sem (enrolled)",
    "Curricular units 1st sem (evaluations)",
    "Curricular units 2nd sem (evaluations)",
    "Curricular units 1st sem (approved)",
    "Curricular units 2nd sem (approved)",
    "Curricular units 1st sem (grade)",
    "Curricular units 2nd sem (grade)",
    "Curricular units 1st sem (without evaluations)",
    "Curricular units 2nd sem (without evaluations)",
]
first_sem_curriculum = {unit for unit in curriculum_units if '1' in unit}
second_sem_curriculum = set(curriculum_units) - first_sem_curriculum

plot_semester_compare_bp(df, curriculum_units)
```



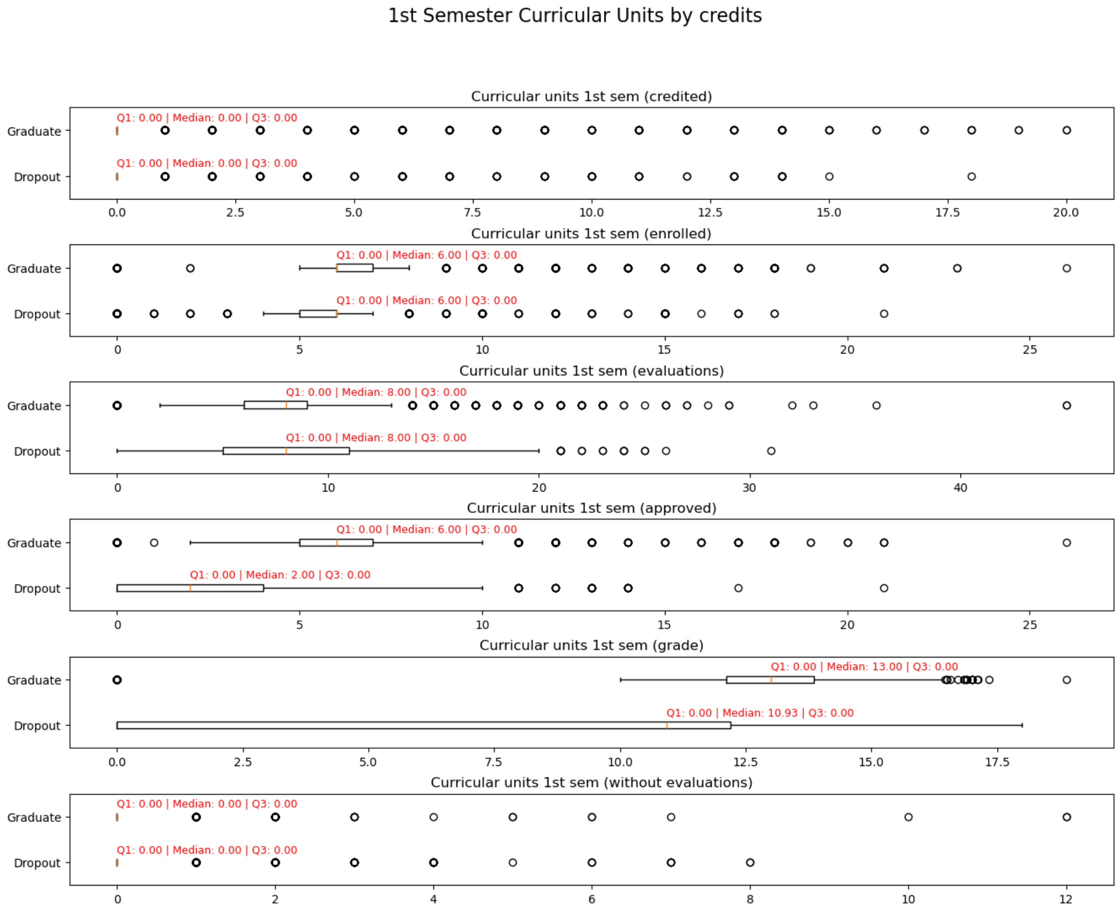
[13]:
```
# will not plot first and semester comparison
df[sorted(list(first_sem_curriculum))].boxplot(vert=False)
plt.title("First semester Curricular Units")
plt.xlabel("Units")
plt.show()
```

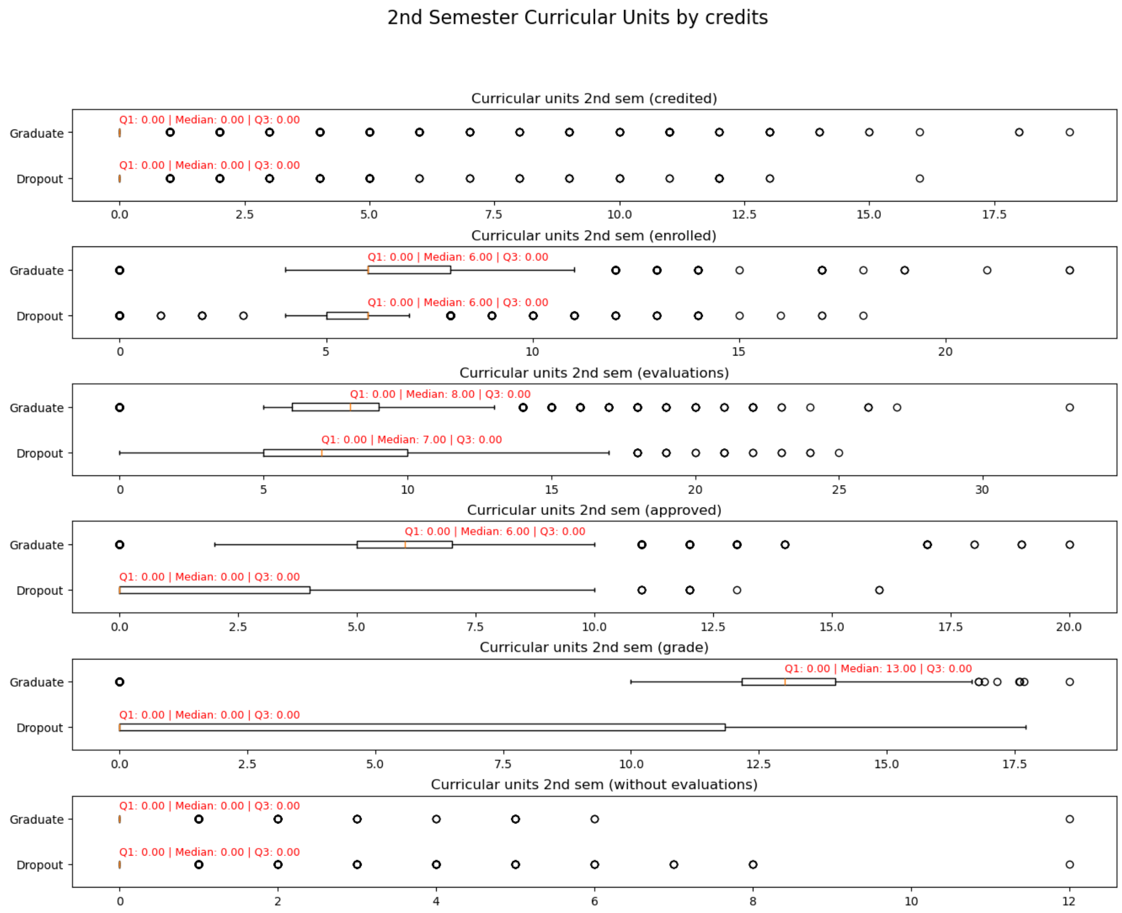
```
df[sorted(list(second_sem_curriculum))].boxplot(vert=False)
plt.title("Second semester Curricular Units")
plt.xlabel("Units")
plt.show()
```



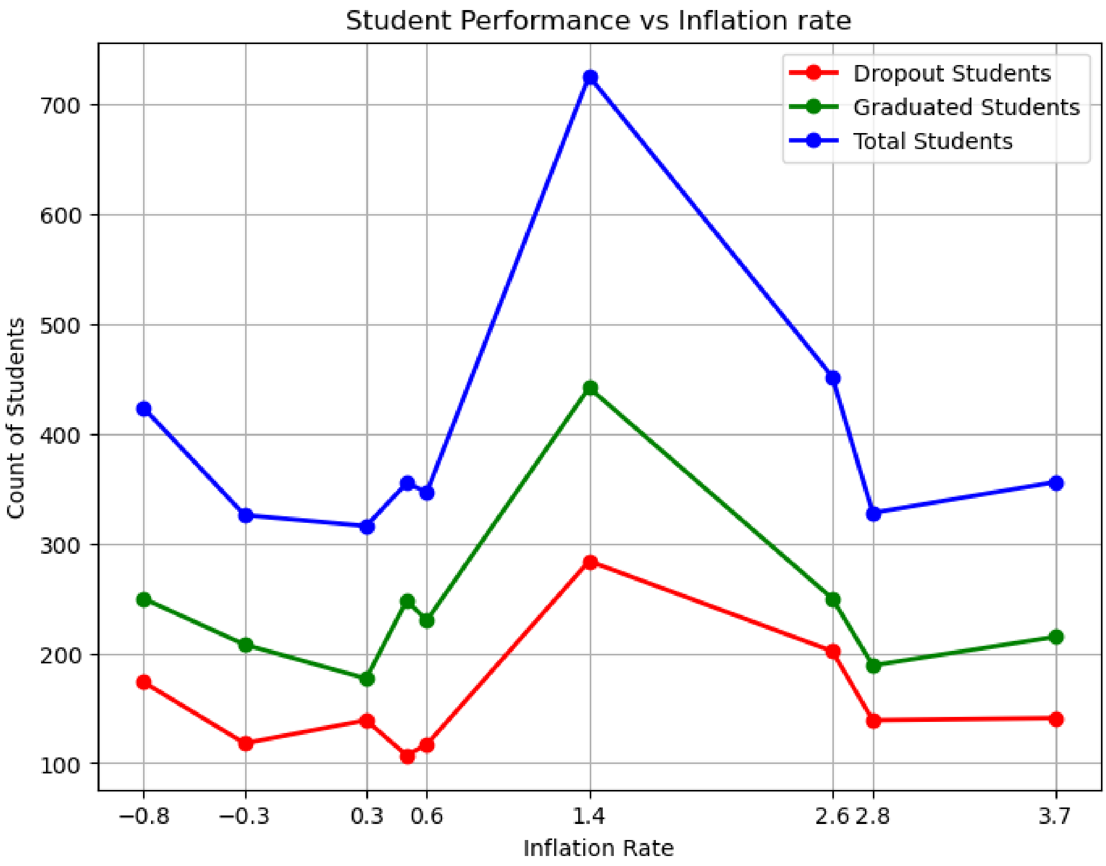


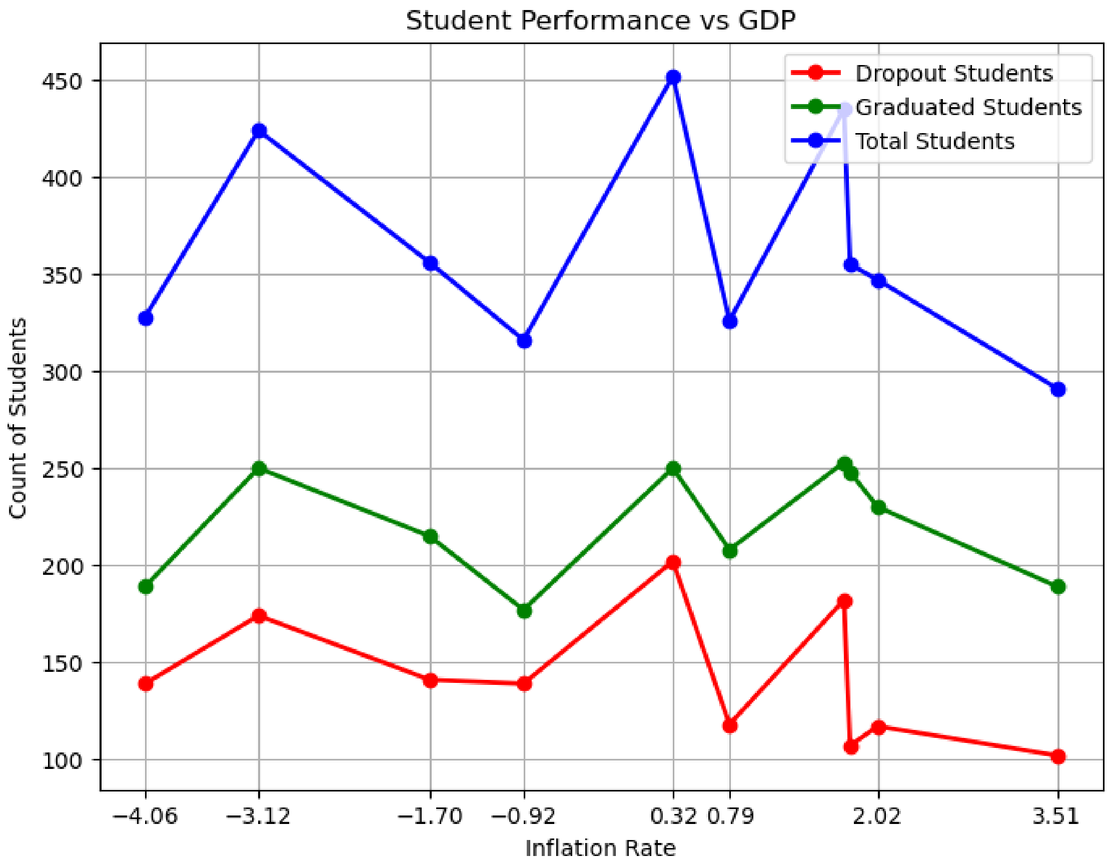
[14]:
```
plot_grad_drop_compare_bp(df, curriculum_units, '1')
plot_grad_drop_compare_bp(df, curriculum_units, '2')
```


1st Semester Curricular Units by credits
Curricular units 1st sem (credited)
Graduate
Dropout
Q1: 0.00 | Median: 0.00 | Q3: 0.00
Q1: 0.00 | Median: 0.00 | Q3: 0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Curricular units 1st sem (enrolled)
Graduate
Dropout
Q1: 0.00 | Median: 6.00 | Q3: 0.00
Q1: 0.00 | Median: 6.00 | Q3: 0.00
0
5
10
15
20
25
Curricular units 1st sem (evaluations)
Graduate
Dropout
Q1: 0.00 | Median: 8.00 | Q3: 0.00
Q1: 0.00 | Median: 8.00 | Q3: 0.00
0
10
20
30
40
Curricular units 1st sem (approved)
Graduate
Dropout
Q1: 0.00 | Median: 6.00 | Q3: 0.00
Q1: 0.00 | Median: 2.00 | Q3: 0.00
0
5
10
15
20
25
Curricular units 1st sem (grade)
Graduate
Dropout
Q1: 0.00 | Median: 13.00 | Q3: 0.00
Q1: 0.00 | Median: 10.93 | Q3: 0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Curricular units 1st sem (without evaluations)
Graduate
Dropout
Q1: 0.00 | Median: 0.00 | Q3: 0.00
Q1: 0.00 | Median: 0.00 | Q3: 0.00
0
2
4
6
8
10
12



```
[15]: plot_line_graph(df, 'Inflation rate')
      plot_line_graph(df, 'GDP')
```


Student Performance vs Inflation rate
Dropout Students
Graduated Students
Total Students
Count of Students
700
600
500
400
300
200
100
Inflation Rate
−0.8
−0.3
0.3
0.6
1.4
2.6
2.8
3.7

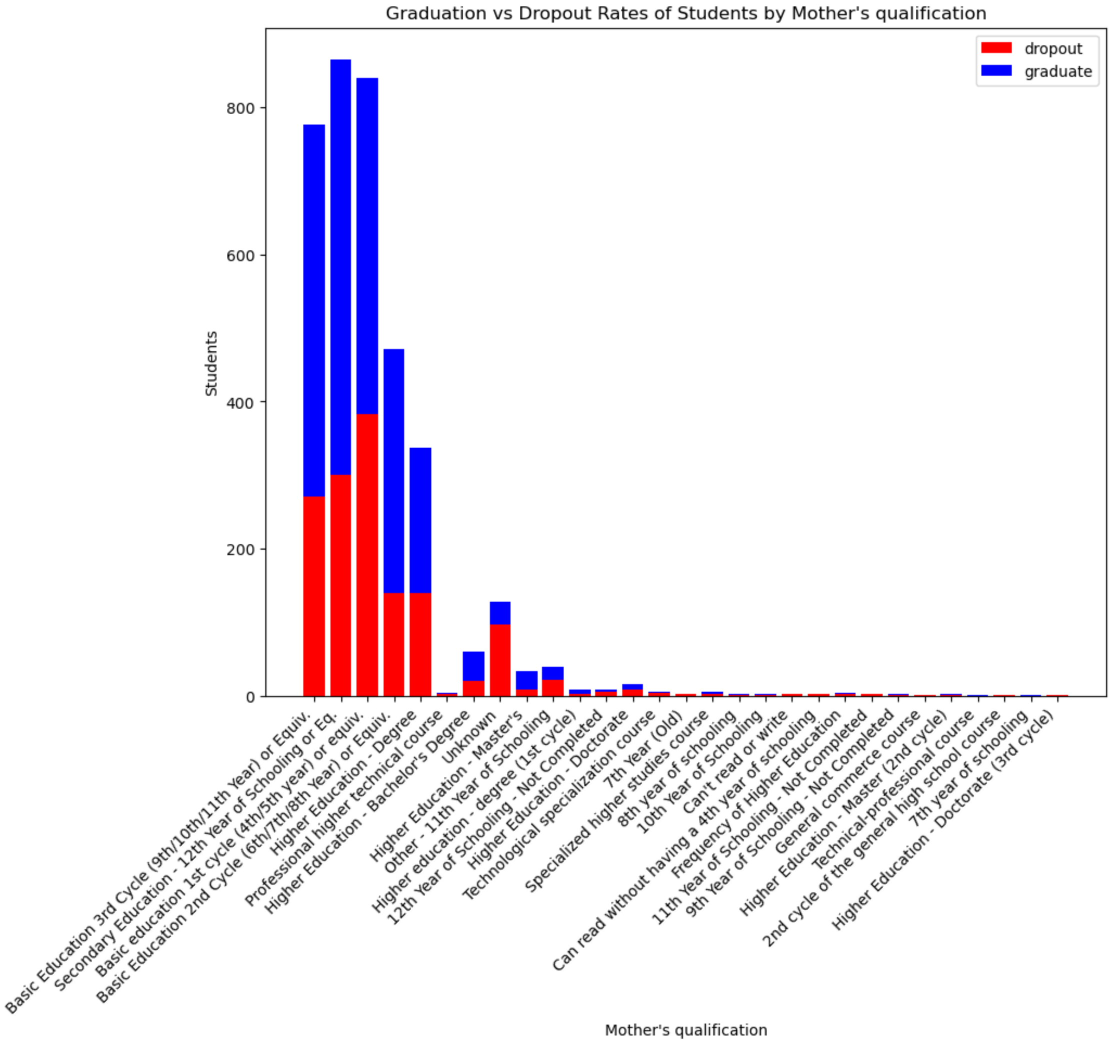


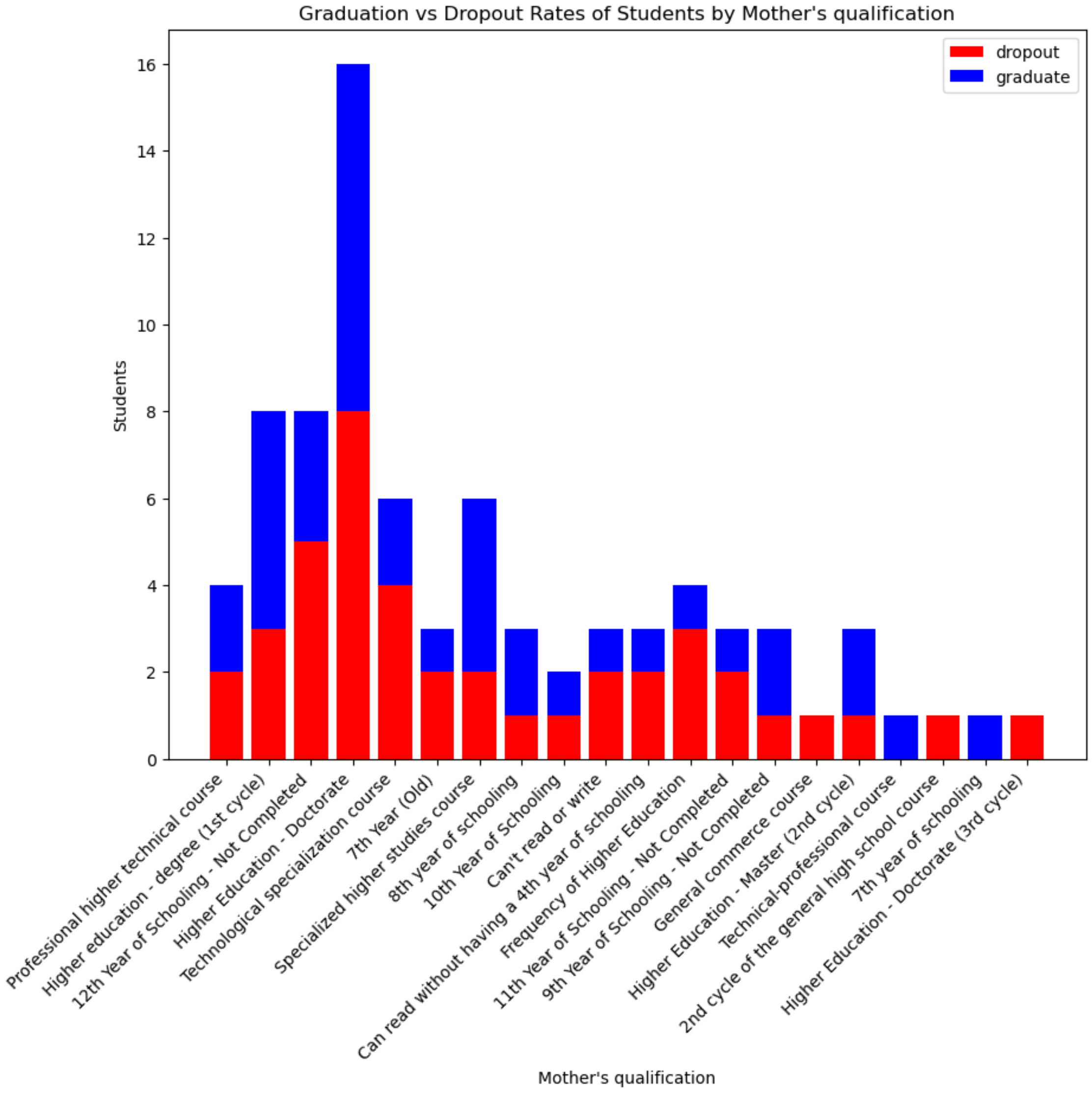
3 Affect of Parental Education on Student Performance

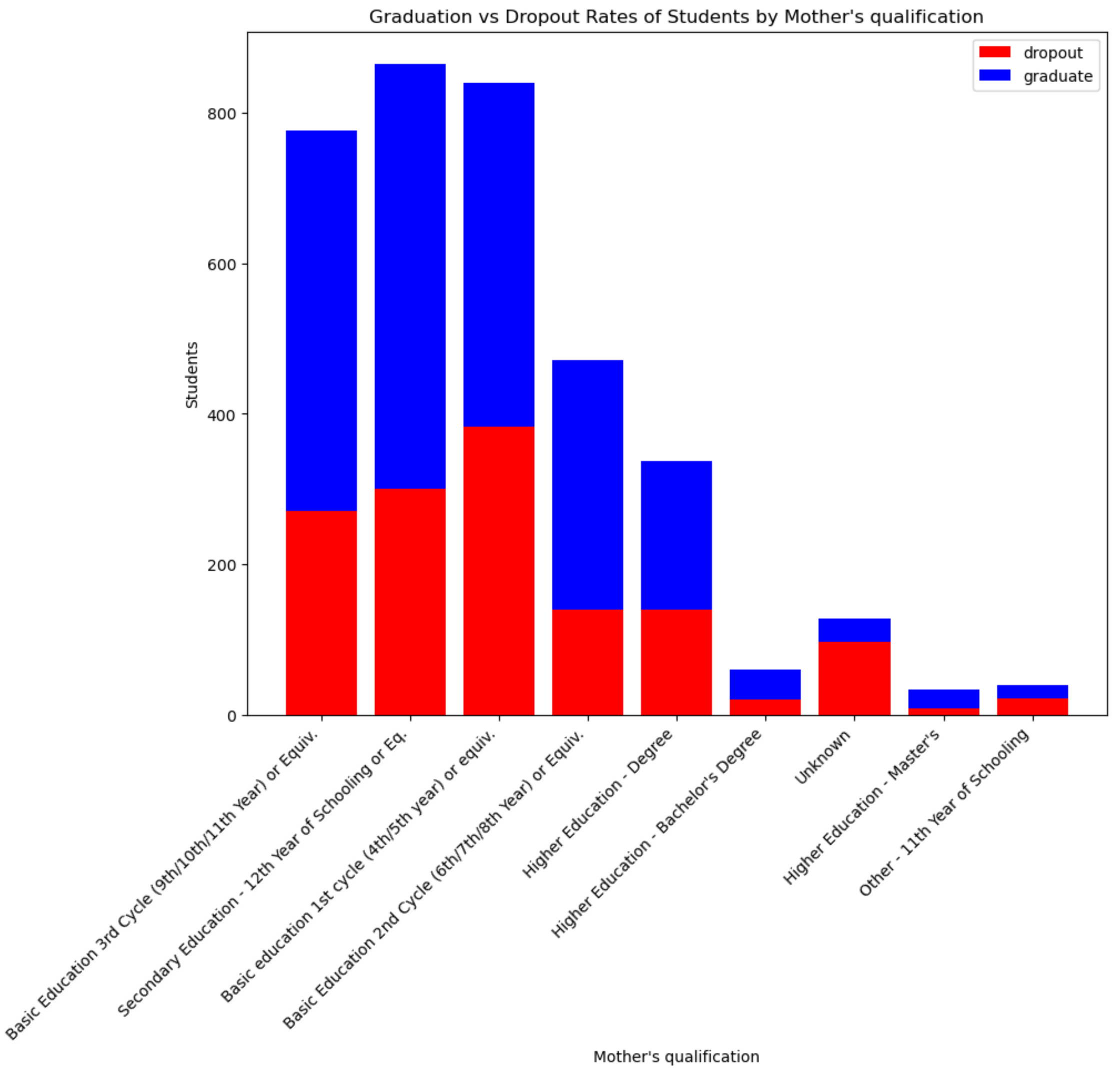
[16]:

```
finished_df = df[df['dropout'].isin(['Dropout', 'Graduate'])]
# filtering out sample categories
mother_quals = finished_df['Mother\'s qualification'].value_counts()
low_mother_quals = list(filter(lambda qual: mother_quals[qual] > 20,␣
  ↪mother_quals.keys()))
high_mother_quals = list(filter(lambda qual: mother_quals[qual] < 20,␣
  ↪mother_quals.keys()))
low_finished_df = finished_df.loc[~(finished_df['Mother\'s qualification'].
  ↪isin(low_mother_quals))]
high_finished_df = finished_df.loc[~(finished_df['Mother\'s qualification'].
  ↪isin(high_mother_quals))]

generate_general_stacked_bar_graph(finished_df, variable_map, 'Mother\'s␣
  ↪qualification')
generate_general_stacked_bar_graph(low_finished_df, variable_map, 'Mother\'s␣
  ↪qualification')
```

```
generate_general_stacked_bar_graph(high_finished_df, variable_map, 'Mother\'s␣
↪qualification')
```

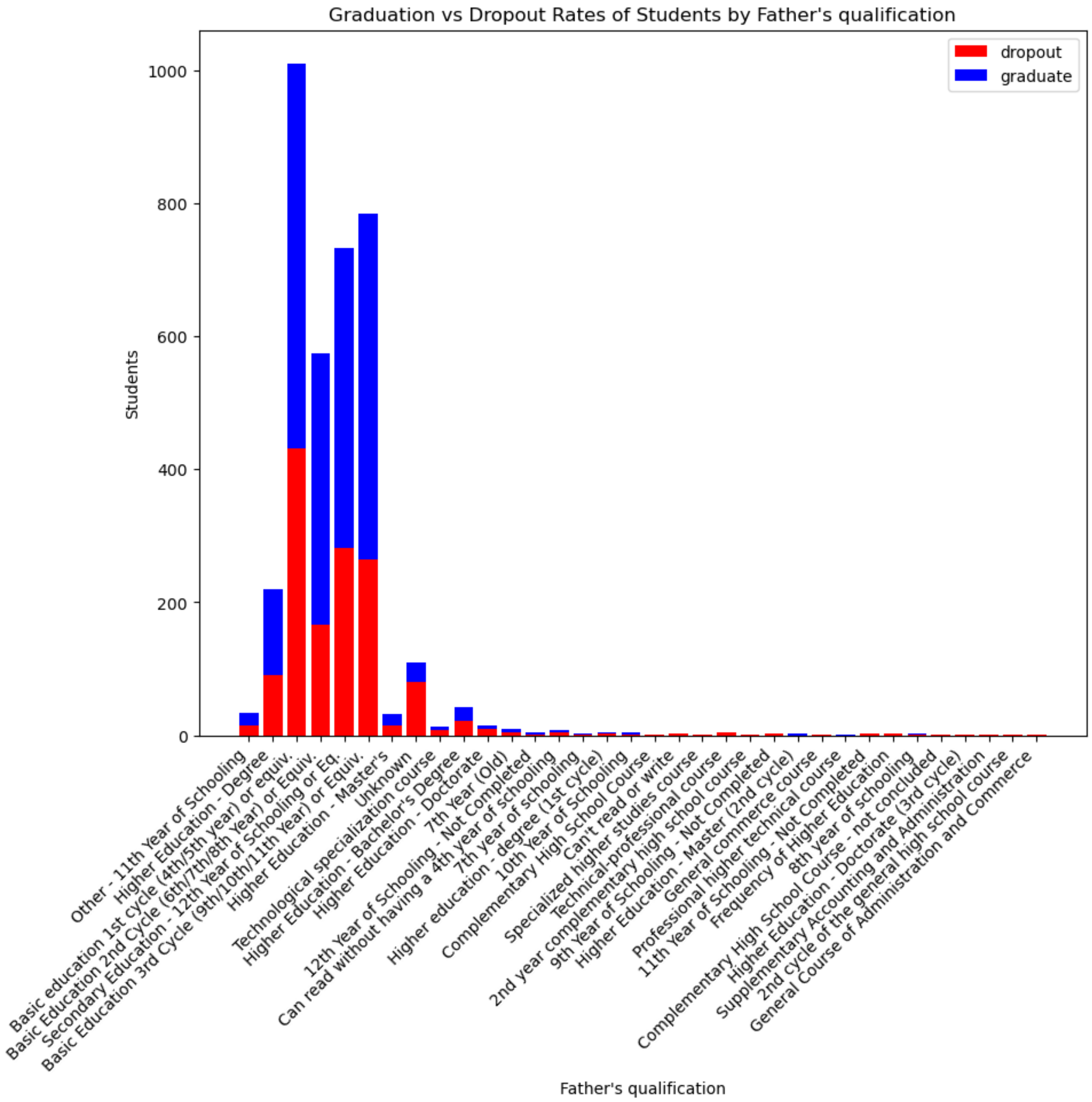


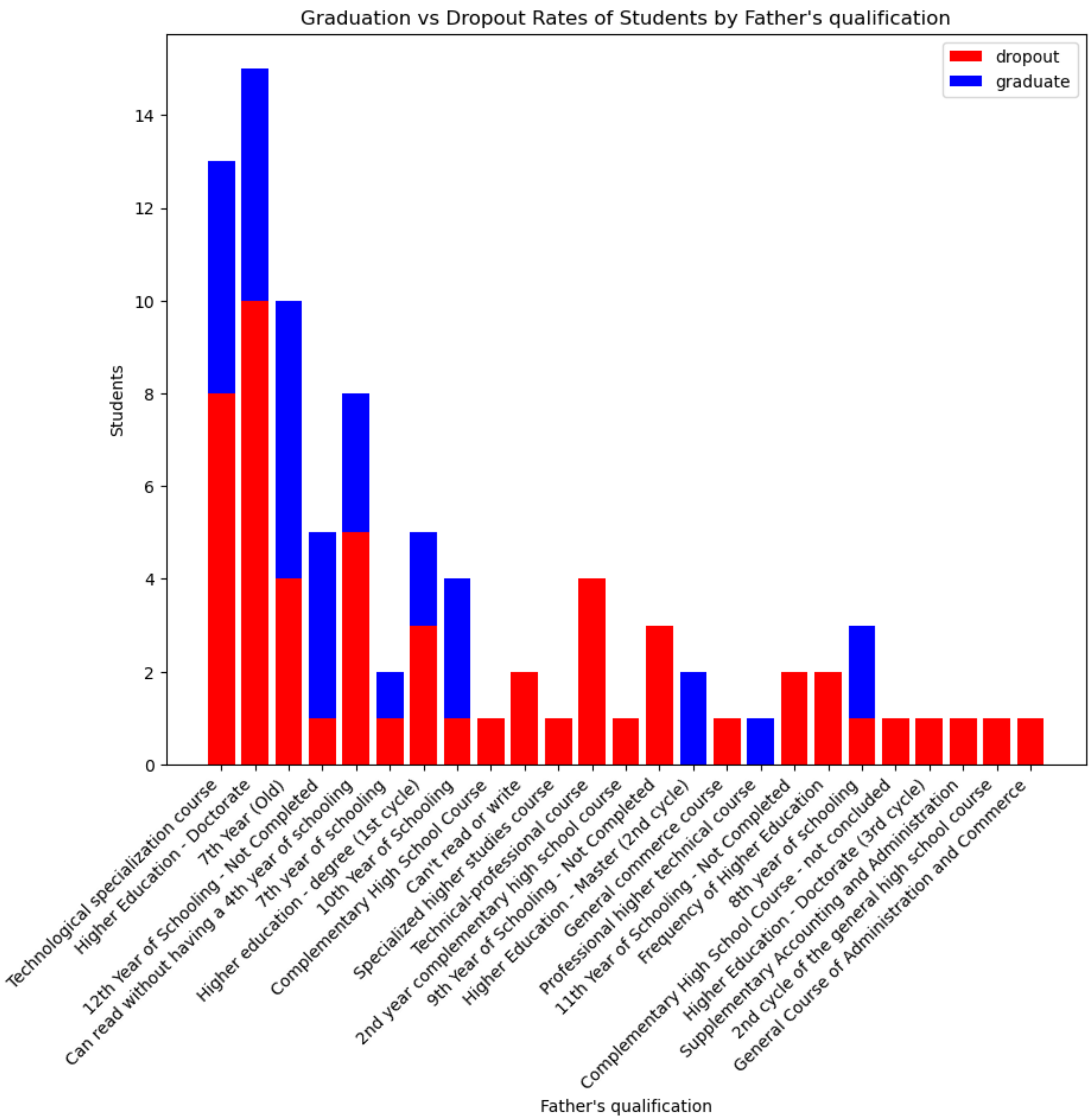

Graduation vs Dropout Rates of Students by Mother's qualification
dropout
graduate
Students
0
2
4
6
8
10
12
14
16
Professional higher technical course
Higher education - degree (1st cycle)
12th Year of Schooling - Not Completed
Higher Education - Doctorate
Technological specialization course
7th Year (Old)
Specialized higher studies course
8th year of schooling
10th Year of Schooling
Can't read or write
Can read without having a 4th year of schooling
Frequency of Higher Education
11th Year of Schooling - Not Completed
9th Year of Schooling - Not Completed
General commerce course
Higher Education - Master (2nd cycle)
Technical-professional course
2nd cycle of the general high school course
7th year of schooling
Higher Education - Doctorate (3rd cycle)
Mother's qualification

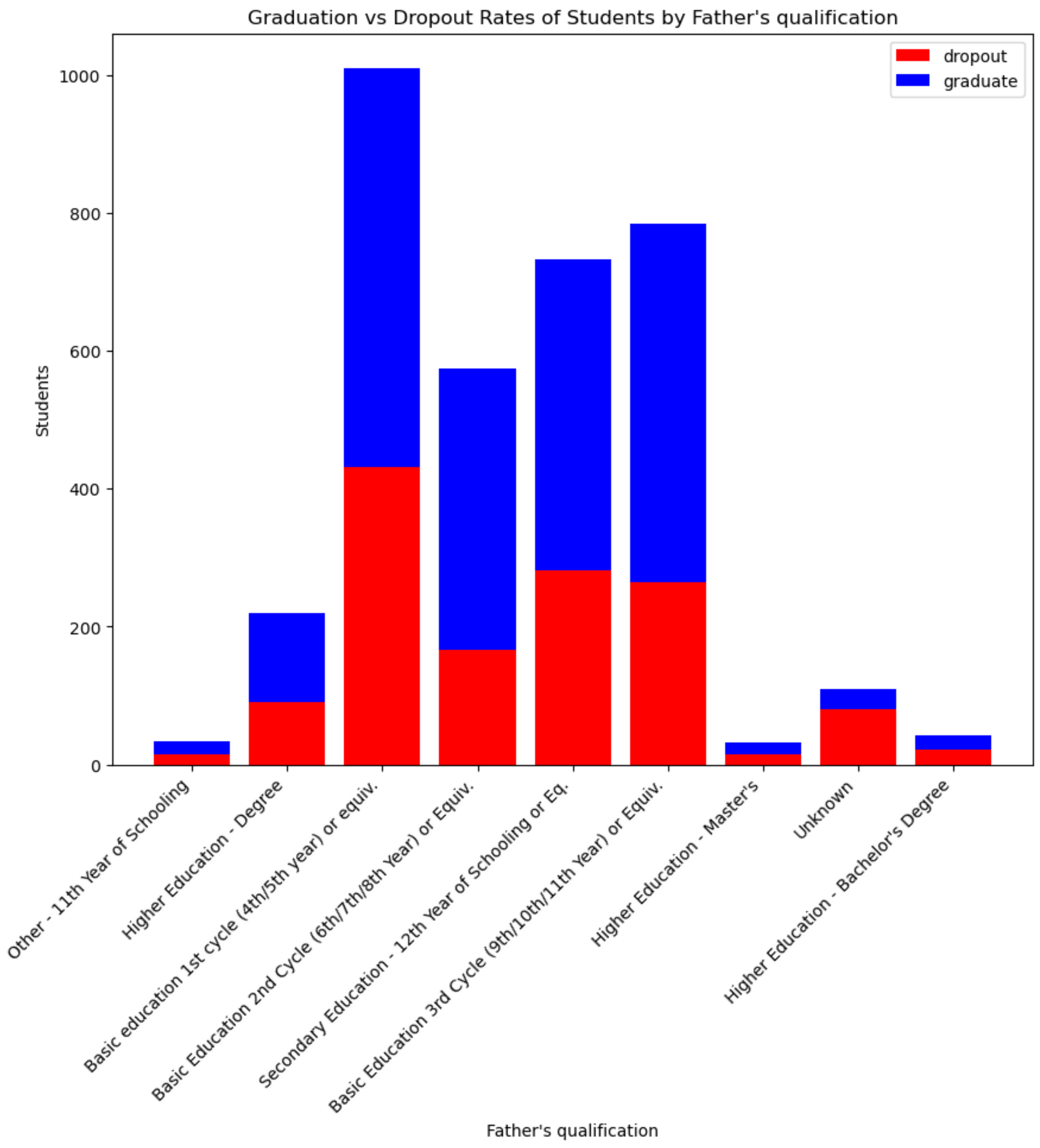


[17]:
```
finished_df = df[df['dropout'].isin(['Dropout', 'Graduate'])]
# filtering out sample categories
father_quals = finished_df['Father\'s qualification'].value_counts()
low_mother_quals = list(filter(lambda qual: father_quals[qual] > 20,␣
 ↪father_quals.keys()))
high_mother_quals = list(filter(lambda qual: father_quals[qual] < 20,␣
 ↪father_quals.keys()))
low_finished_df = finished_df.loc[~(finished_df['Father\'s qualification'].
 ↪isin(low_mother_quals))]
high_finished_df = finished_df.loc[~(finished_df['Father\'s qualification'].
 ↪isin(high_mother_quals))]

generate_general_stacked_bar_graph(finished_df, variable_map, 'Father\'s␣
 ↪qualification')
```

```
generate_general_stacked_bar_graph(low_finished_df, variable_map, 'Father\'s␣
↪qualification')
generate_general_stacked_bar_graph(high_finished_df, variable_map, 'Father\'s␣
↪qualification')
```




Graduation vs Dropout Rates of Students by Father's qualification
🟥 dropout
🟦 graduate
Students
0
2
4
6
8
10
12
14
Technological specialization course
Higher Education - Doctorate
7th Year (Old)
12th Year of Schooling - Not Completed
Can read without having a 4th year of schooling
7th year of schooling
Higher education - degree (1st cycle)
10th Year of Schooling
Complementary High School Course
Can't read or write
Specialized higher studies course
Technical-professional course
2nd year complementary high school course
9th Year of Schooling - Not Completed
Higher Education - Master (2nd cycle)
General commerce course
Professional higher technical course
11th Year of Schooling - Not Completed
Frequency of Higher Education
8th year of schooling
Complementary High School Course - not concluded
Higher Education - Doctorate (3rd cycle)
Supplementary Accounting and Administration
2nd cycle of the general high school course
General Course of Administration and Commerce
Father's qualification


Graduation vs Dropout Rates of Students by Father's qualification
🟥 dropout
🟦 graduate
Students
0
200
400
600
800
1000
Other - 11th Year of Schooling
Higher Education - Degree
Basic education 1st cycle (4th/5th year) or equiv.
Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.
Secondary Education - 12th Year of Schooling or Eq.
Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.
Higher Education - Master's
Unknown
Higher Education - Bachelor's Degree
Father's qualification

```
[17]:
```