EXPLORATORY DATA ANALYSIS

OCHALO DEOGRATIUS

2023/HD05/20127U

2300720127

Introduction

The urgency to predict the academic performance of students in many educational institutions arose when it was noted that the rate at which students are dropping out of school was growing very fast. This led to many nations devising and researching on methods they can use to retain students in schools. The Exploratory Data Analysis I will be doing is for a dataset of one of the many that researchers and scholars have tackled. Predict Students' Dropout and Academic Success is a dataset that was designed with purpose of providing formulas to predict the level of academic retentions in higher institutions.

It is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course**.** It consists of 4424 rows and 37 columns. In my analysis I will be looking at the Factors Causing Students' Dropout and Academic Success. How academic performance has an effect on the Students' dropout and academic success? Does unemployment rate contribute to Students' dropout and academic success? What does Father's qualification have do with Students' dropout and academic success?

This dataset contains demographics, social-economic factors, and academic performance variables related to undergraduate students that will be used to investigate the impact on student dropout and academic success. Below are the feature descriptions of the dataset.

- **Marital status:** The marital status of the student. (Categorical)- 1 – single 2 – married 3 – widower 4 – divorced 5 – facto union 6 – legally separated
- **Application mode**: The method of application used by the student. (Categorical)- 1 - 1st phase - general contingent 2 - Ordinance No. 612/93 5 - 1st phase - special contingent (Azores Island) 7 - Holders of other higher courses 10 - Ordinance No. 854-B/99 15 - International student (bachelor) 16 - 1st phase - special contingent (Madeira Island) 17 - 2nd phase - general contingent 18 - 3rd phase - general contingent 26 - Ordinance No. 533-A/99, item b2) (Different Plan) 27 - Ordinance No. 533-A/99, item b3 (Other Institution) 39 - Over 23 years old 42 - Transfer 43 - Change of course 44 - Technological specialization diploma holders 51 - Change of institution/course 53 - Short cycle diploma holders 57 - Change of institution/course (International)
- **Application order:** The order in which the student applied. (Numerical)- Application order (between 0 - first choice; and 9 last choice)
- **Course:** The course taken by the student. (Categorical)- 33 - Biofuel Production Technologies 171 - Animation and Multimedia Design 8014 - Social Service (evening attendance) 9003 - Agronomy 9070 - Communication Design 9085 - Veterinary Nursing 9119 - Informatics Engineering 9130 - Equinculture 9147 - Management 9238 - Social Service 9254 - Tourism 9500 - Nursing 9556 - Oral Hygiene 9670 - Advertising

and Marketing Management 9773 - Journalism and Communication 9853 - Basic Education 9991 - Management (evening attendance)

- **Daytime/evening attendance**: Whether the student attends classes during the day or in the evening. (Categorical)- 1 – daytime 0 - evening
- **Previous qualification:** The qualification obtained by the student before enrolling in higher education. (Categorical)- 1 - Secondary education 2 - Higher education - bachelor's degree 3 - Higher education - degree 4 - Higher education - master's 5 - Higher education - doctorate 6 - Frequency of higher education 9 - 12th year of schooling - not completed 10 - 11th year of schooling - not completed 12 - Other - 11th year of schooling 14 - 10th year of schooling 15 - 10th year of schooling - not completed 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv. 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 42 - Professional higher technical course 43 - Higher education - master (2nd cycle)
- **Previous qualification (grade):** Grade of previous qualification (between 0 and 200)
- **Nationality**: The nationality of the student. (Categorical)- 1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26 - Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldova (Republic of); 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 - Colombian
- **Mother's qualification:** The qualification of the student's mother. (Categorical) - 1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) 12 - Other - 11th Year of Schooling 14 - 10th Year of Schooling 18 - General commerce course 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv. 22 - Technical-professional course 26 - 7th year of schooling 27 - 2nd cycle of the general high school course 29 - 9th Year of Schooling - Not Completed 30 - 8th year of schooling 34 - Unknown 35 - Can't read or write 36 - Can read without having a 4th year of schooling 37 - Basic education 1st cycle (4th/5th year) or equiv. 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 41 - Specialized higher studies course 42 - Professional higher technical course 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)
- **Father's qualification:** The qualification of the student's father. (Categorical)- 1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) 12 - Other - 11th Year of Schooling 13 - 2nd year complementary high school course 14 - 10th Year of Schooling 18 - General commerce course 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv. 20 - Complementary High School Course 22 -

Technical-professional course 25 - Complementary High School Course - not concluded 26 - 7th year of schooling 27 - 2nd cycle of the general high school course 29 - 9th Year of Schooling - Not Completed 30 - 8th year of schooling 31 - General Course of Administration and Commerce 33 - Supplementary Accounting and Administration 34 - Unknown 35 - Can't read or write 36 - Can read without having a 4th year of schooling 37 - Basic education 1st cycle (4th/5th year) or equiv. 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 41 - Specialized higher studies course 42 - Professional higher technical course 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)

Mother's occupation: The occupation of the student's mother. (Categorical)- 0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers 10 - Armed Forces Professions 90 - Other Situation 99 - (blank) 122 - Health professionals 123 - teachers 125 - Specialists in information and communication technologies (ICT) 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 171 - Skilled construction workers and the like, except electricians 173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like 175 - Workers in food processing, woodworking, clothing and other industries and crafts 191 - cleaning workers 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants

Father's occupation: The occupation of the student's father. (Categorical)- 0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers 10 - Armed Forces Professions 90 - Other Situation 99 - (blank) 101 - Armed Forces Officers 102 - Armed Forces Sergeants 103 - Other Armed Forces personnel 112 - Directors of administrative and commercial services 114 - Hotel, catering, trade and other services directors 121 - Specialists in the physical sciences, mathematics, engineering and

related techniques 122 - Health professionals 123 - teachers 124 - Specialists in finance, accounting, administrative organization, public and commercial relations 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services 135 - Information and communication technology technicians 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 154 - Protection and security services personnel 161 - Market-oriented farmers and skilled agricultural and animal production workers 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence 171 - Skilled construction workers and the like, except electricians 172 - Skilled workers in metallurgy, metalworking and similar 174 - Skilled workers in electricity and electronics 175 - Workers in food processing, woodworking, clothing and other industries and crafts 181 - Fixed plant and machine operators 182 - assembly workers 183 - Vehicle drivers and mobile equipment operators 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants 195 - Street vendors (except food) and street service providers

- **Displaced:** Whether the student is a displaced person. (Categorical)- 1 – yes 0 – no
- **Educational special needs:** Whether the student has any special educational needs (Categorical)- 1 – yes 0 – no
- **Debtor:** Whether the student is a debtor. (Categorical)- 1 – yes 0 – no
- **Tuition fees up to date:** Whether the student's tuition fees are up to date. (Categorical)
- **Gender:** The gender of the student. (Categorical)- 1 – male 0 – female
- **Scholarship holder:** Whether the student is a scholarship holder (Categorical)- 1 – yes 0 – no
- **Age at enrollment:** The age of the student at the time of enrollment. (Numerical)
- **International:** Whether the student is an international student. (Categorical)- 1 – yes 0 – no
- **Curricular units 1st sem (credited**): The number of curricular units credited by the student in the first semester. (Numerical)- Number of curricular units credited in the 1st semester
- **Curricular units 1st sem (enrolled):** The number of curricular units enrolled by the student in the first semester. (Numerical)- Number of curricular units enrolled in the 1st semester
- **Curricular units 1st sem (evaluations):** The number of curricular units evaluated by the student in the first semester. (Numerical)- Number of evaluations to curricular units in the 1st semester
- **Curricular units 1st sem (approved):** The number of curricular units approved by the student in the first semester. (Numerical)- Number of curricular units approved in the 1st semester

- **Target:** whether the student dropped out or graduated.

We are going first to understand the statistics of the dataset. We use data_frame.describe() to get the details.

```
In [63]: # Getting the overall statistics of our Dataset

In [138]: data_frame.describe()
```

Out[138]:

	Marital status	Application mode	Application order	Course	Daytime/evening attendance\t	Previous qualification	Previous qualification (grade)	Nationality	Mother's qualification	Father's qualification	...	Curricular units 1st sem (without evaluations)
count	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	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	132.613314	1.873192	19.561935	22.275316	...	0.137658
std	0.605747	17.484682	1.313793	2063.566416	0.311897	10.216592	13.188332	6.914514	15.603186	15.343108	...	0.690880
min	1.000000	1.000000	0.000000	33.000000	0.000000	1.000000	95.000000	1.000000	1.000000	1.000000	...	0.000000
25%	1.000000	1.000000	1.000000	9085.000000	1.000000	1.000000	125.000000	1.000000	2.000000	3.000000	...	0.000000
50%	1.000000	17.000000	1.000000	9238.000000	1.000000	1.000000	133.100000	1.000000	19.000000	19.000000	...	0.000000
75%	1.000000	39.000000	2.000000	9556.000000	1.000000	1.000000	140.000000	1.000000	37.000000	37.000000	...	0.000000
max	6.000000	57.000000	9.000000	9991.000000	1.000000	43.000000	190.000000	109.000000	44.000000	44.000000	...	12.000000

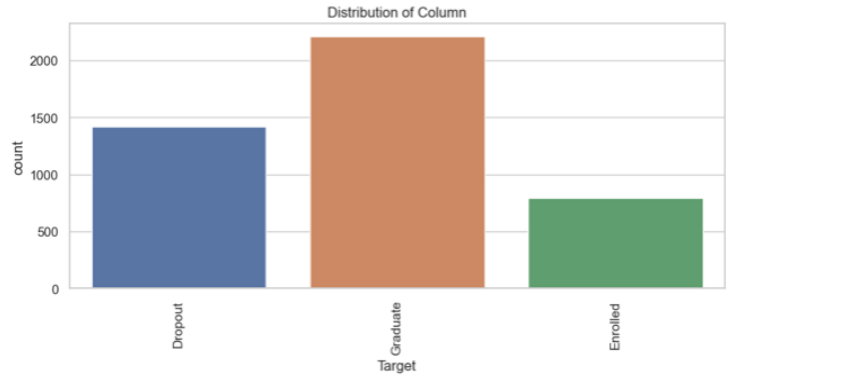
8 rows × 36 columns

```
In [ ]:
```

In this data frame, we have 5 quantitative features that have been randomly selected to be used in our data analysis (previous qualification (grade), father's qualification, admission grade, age at enrollment and unemployment rate)

EXPLORATORY DATA ANALYSIS

```
In [518]: categorical_columns = data_frame.select_dtypes(include=['object']).columns
          for column in categorical_columns:
              plt.figure(figsize=(10, 4))
              sns.countplot(x=column, data=data_frame)
              plt.title('Distribution of Column')
              plt.xticks(rotation=90)
              plt.show()
```



The data shows that the number of School dropouts, Graduates and the Students enrolled is not balanced. The number of Graduates is having the highest number of students followed by the number of students dropout and then the Students enrolled has the lowest number.

UNIVARIATE ANALYSIS:

The histogram plot provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values called bins.

```
In [ ]: # Histogram plots showing the Frequency distribution of Student Dropout and Academic success(SDAS) by:
        #Previous Qualification grades
        #Father's qualification
        #Admission grade
        #Age at enrollment
        #Unemployment rate
```

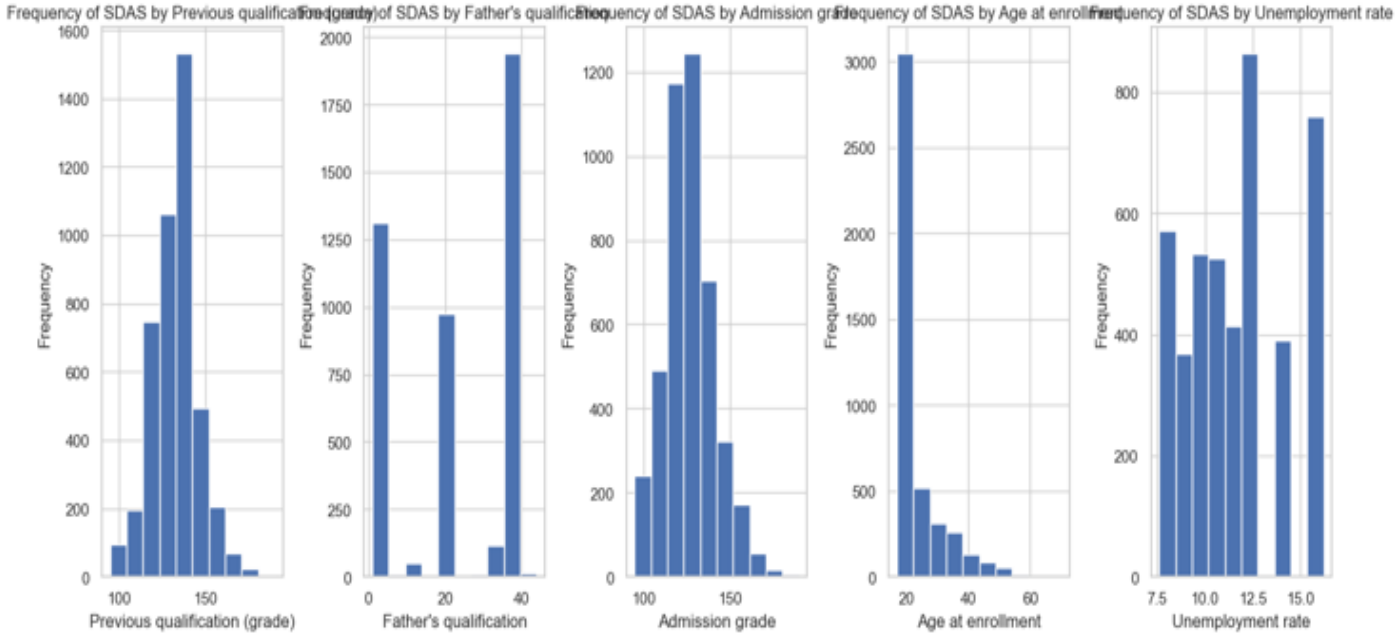
```
In [198]: plt.figure(figsize=(14,6))
          plt.subplot(1,5,1)
          data_frame['Previous qualification (grade)'].plot(kind='hist')
          plt.title('Frequency of SDAS by Previous qualification (grade)')
          plt.xlabel('Previous qualification (grade)')
          plt.tight_layout()


          plt.subplot(1,5,2)
          data_frame["Father's qualification"].plot(kind='hist')
          plt.title("Frequency of SDAS by Father's qualification")
          plt.xlabel("Father's qualification")
          plt.tight_layout()


          plt.subplot(1,5,3)
          data_frame['Admission grade'].plot(kind='hist')
          plt.title('Frequency of SDAS by Admission grade')
          plt.xlabel('Admission grade')
          plt.tight_layout()


          plt.subplot(1,5,4)
          data_frame['Age at enrollment'].plot(kind='hist')
          plt.title('Frequency of SDAS by Age at enrollment')
          plt.xlabel('Age at enrollment')
          plt.tight_layout()


          plt.subplot(1,5,5)
          data_frame['Unemployment rate'].plot(kind='hist')
          plt.title('Frequency of SDAS by Unemployment rate')
          plt.xlabel('Unemployment rate')
          plt.tight_layout()
```



In [ ]:
```
#Box plot to identify outliers
#A plot box of the statistics Students Dropout and Academic Success (SDAS)by:
#Previous Qualification grades
#Father's qualification
#Admission grade
#Age at enrollment
#Unemployment rate
```

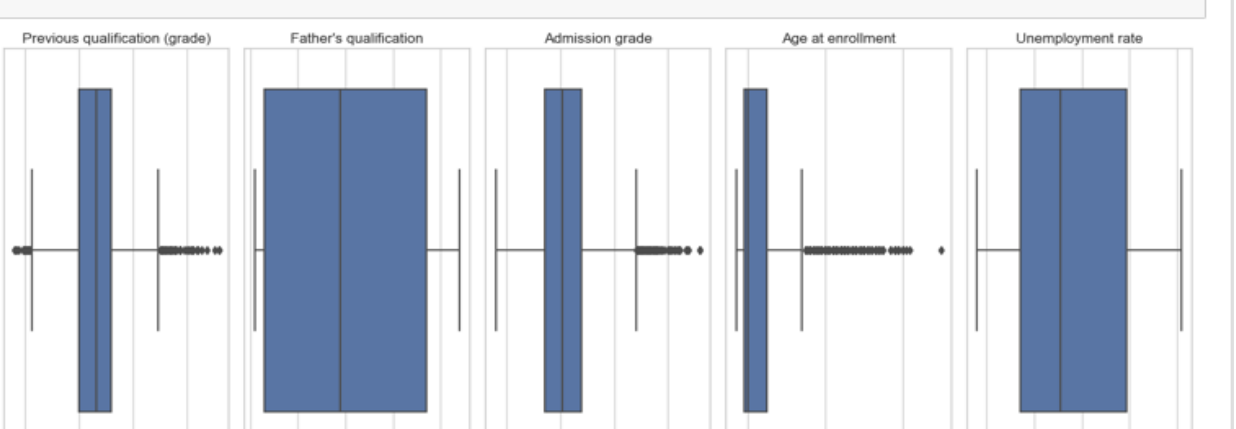
In [158]:
```
plt.figure(figsize=(14,6))
plt.subplot(1,5,1)
sns.boxplot(data=data_frame,x='Previous qualification (grade)')
plt.title('Previous qualification (grade)')
plt.tight_layout()


plt.subplot(1,5,2)
sns.boxplot(data=data_frame,x="Father's qualification")
plt.title("Father's qualification")
plt.tight_layout()


plt.subplot(1,5,3)
sns.boxplot(data=data_frame,x='Admission grade')
plt.title('Admission grade')
plt.tight_layout()


plt.subplot(1,5,4)
sns.boxplot(data=data_frame,x='Age at enrollment')
plt.title('Age at enrollment')
plt.tight_layout()


plt.subplot(1,5,5)
sns.boxplot(data=data_frame,x='Unemployment rate')
plt.title('Unemployment rate')
plt.tight_layout()
```

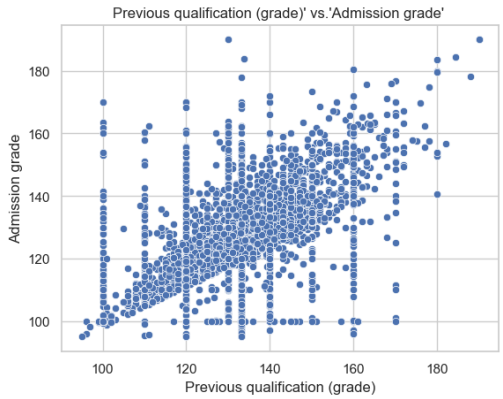


The above box plot show that the dataset has outliers in previous qualification(grade), admission grade, and on the age at enrollment data. This calls for further analysis and training of the dataset

BIVARIATE ANALYSIS

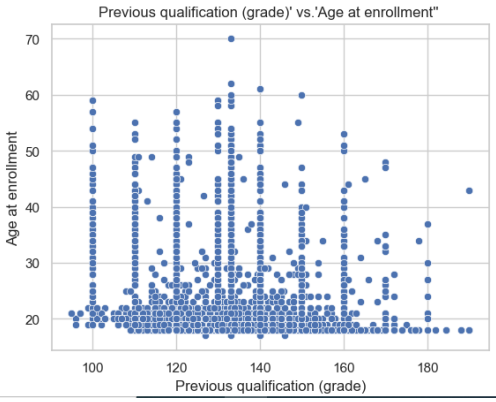
#Scatter plot comparing the different features

```
In [241]: sns.scatterplot(data_frame['Previous qualification (grade)'],data['Admission grade'])
          plt.title("Previous qualification (grade)' vs.'Admission grade' ")
Out[241]: Text(0.5, 1.0, "Previous qualification (grade)' vs.'Admission grade' ")
```



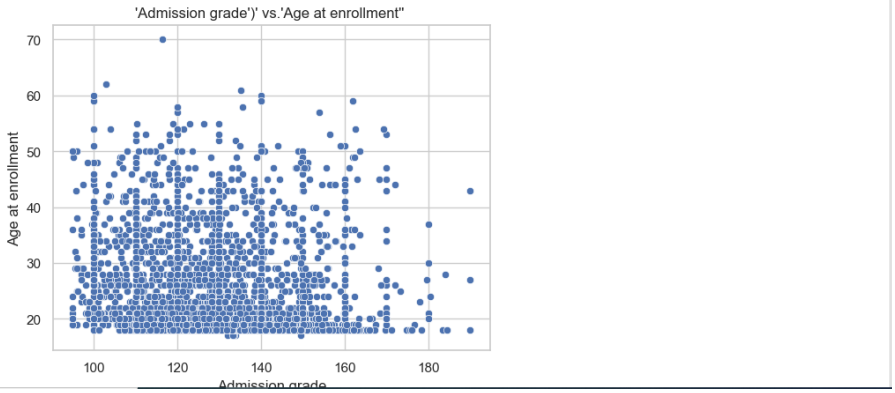
The figure above shows that there is a very strong positive linear relationship between admission grade and previous qualification(grade). It shows that the students of Admission grade of 100 had lower previous qualification(grades). The data also shows that an increase in the Admission grade leads to increase in Previous qualification(grade)

```
In [243]: sns.scatterplot(data_frame['Previous qualification (grade)'],data_frame['Age at enrollment'])
          plt.title("Previous qualification (grade)' vs.'Age at enrollment'' ")
Out[243]: Text(0.5, 1.0, "Previous qualification (grade)' vs.'Age at enrollment'' ")
```



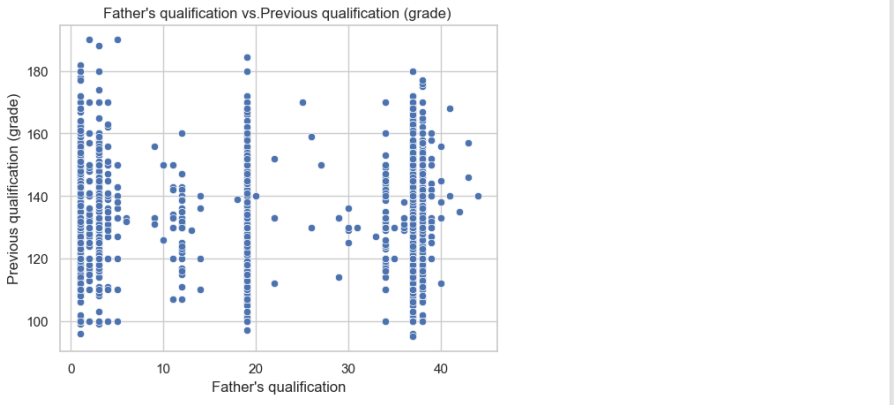
The figure above shows that there is a very strong relationship between age at enrollment and previous qualification(grade) from the age of 17 upto around the age of 27.

```
In [244]: sns.scatterplot(data_frame['Admission grade'],data_frame['Age at enrollment'])
          plt.title("'Admission grade')' vs.'Age at enrollment'' ")
Out[244]: Text(0.5, 1.0, "'Admission grade')' vs.'Age at enrollment'' ")
```



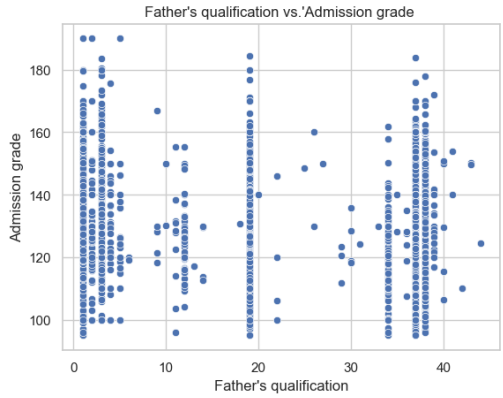
The figure above shows that there is a very strong relationship between Age at enrollment and Admission grade from the age of 17 to the age of 45.

```
In [245]: sns.scatterplot(data_frame["Father's qualification"],data_frame['Previous qualification (grade)'])
          plt.title("Father's qualification vs.Previous qualification (grade) ")
Out[245]: Text(0.5, 1.0, "Father's qualification vs.Previous qualification (grade) ")
```

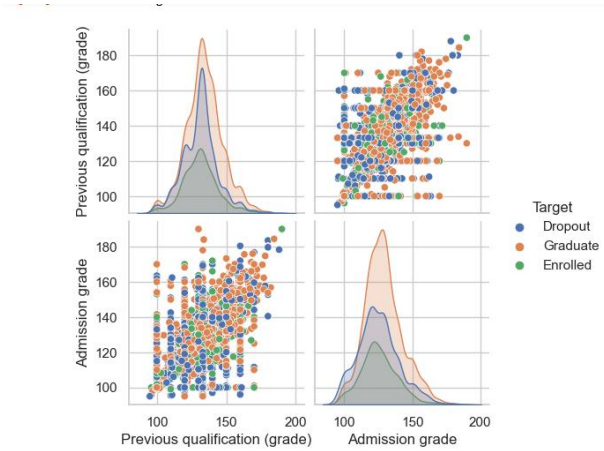


The figure above shows that there is moderate relationship between Previous Qualification(grade) and Father’s qualication.

```
In [246]: sns.scatterplot(data_frame["Father's qualification"],data_frame[ 'Admission grade'])
          plt.title("Father's qualification vs.'Admission grade ")
Out[246]: Text(0.5, 1.0, "Father's qualification vs.'Admission grade ")
```



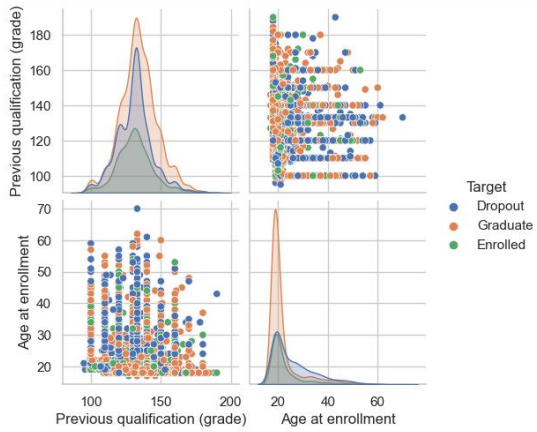
The figure above shows that there is a moderate relationship between Admission grade and Father's Qualification.

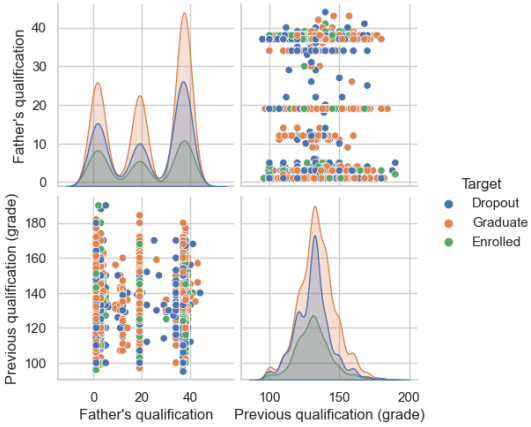


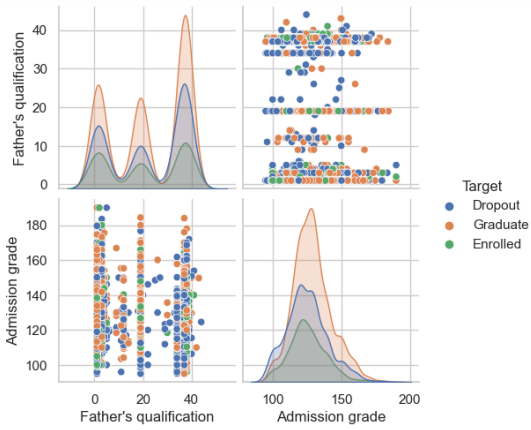
Graduate points with high admission grades tend to also have high previous qualification (grade), indicating a very strong positive correlation between admission grades and previous qualification (grade).

Dropout Points tend to increase from administration grade as the previous qualifications grade also increase. This indicates that the admission grade has a strong negative relationship with the previous qualifications grade.

The enroll points are spread out but tend to be in the mid-range of grades, indicating that these students are still progressing but may not have exceptionally high grades.


Previous qualification (grade)
Age at enrollment
Previous qualification (grade)
Age at enrollment
Target
Dropout
Graduate
Enrolled


Father's qualification
Previous qualification (grade)
Father's qualification
Previous qualification (grade)
Target
Dropout
Graduate
Enrolled

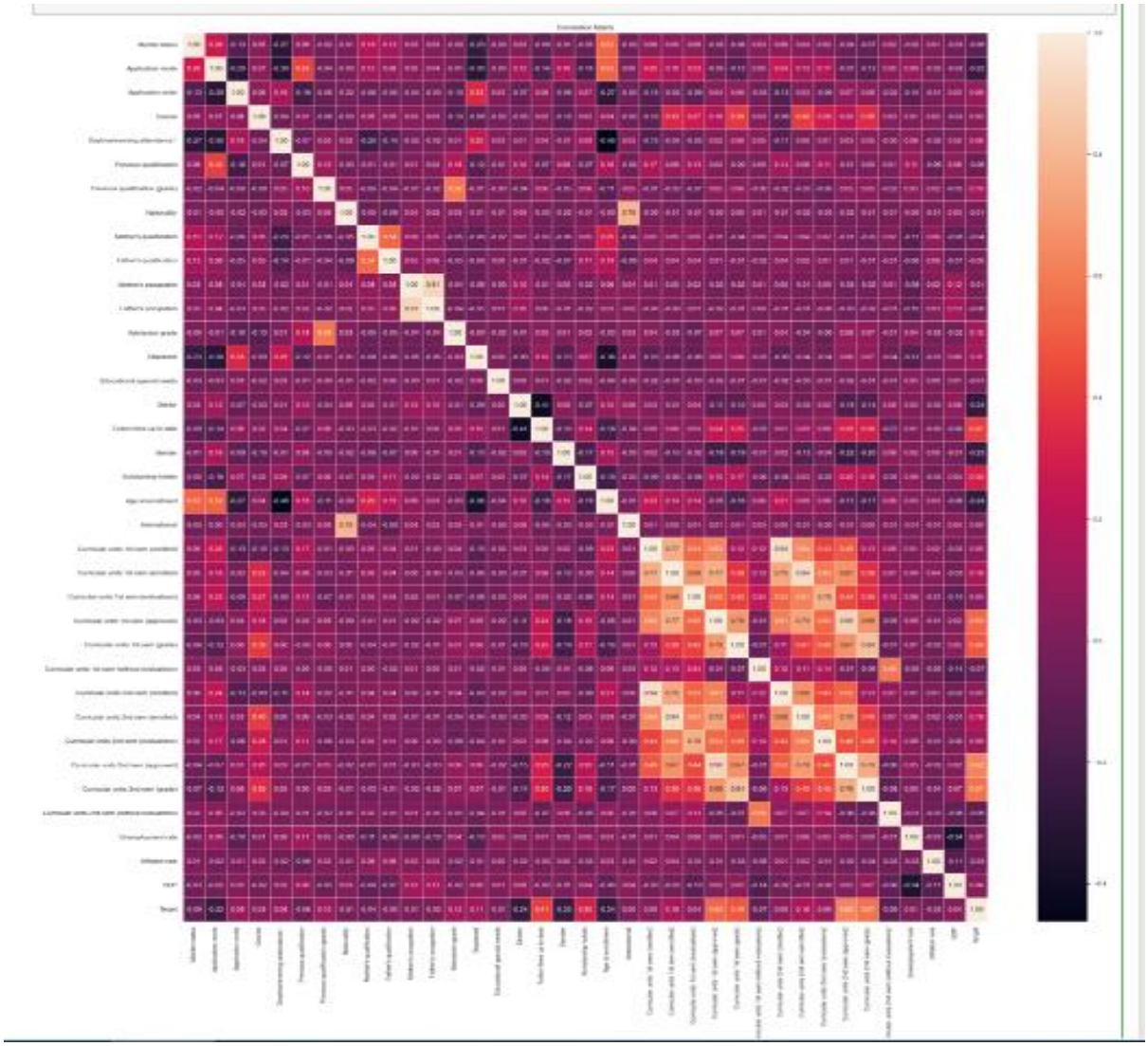

Father's qualification
Admission grade
Father's qualification
Admission grade
Target
Dropout
Graduate
Enrolled

In [ ]:
```
#Finding out which features are correlated with Target
```

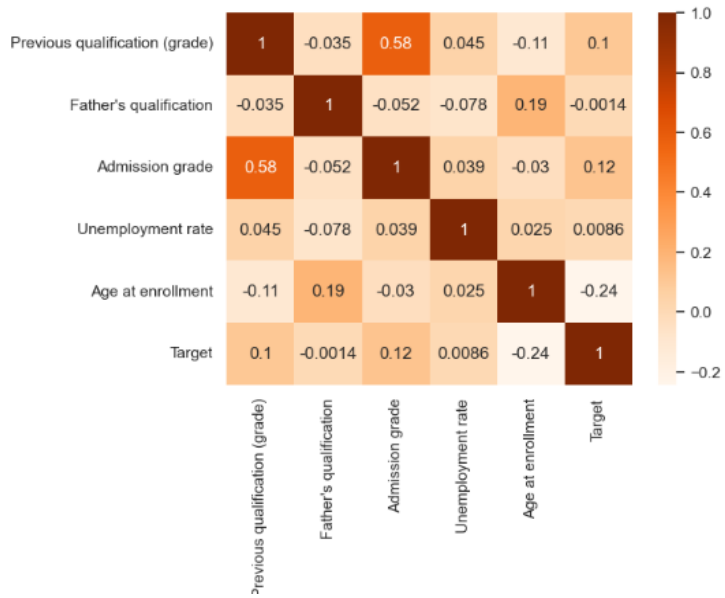
In [251]:
```
label_encoder = LabelEncoder()
data_frame['Target'] = label_encoder.fit_transform(data_frame['Target'])
```

In [252]:
```
correlation_matrix = data_frame.corr()

plt.figure(figsize=(30, 30))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```



```
In [264]: Father's qualification",'Admission grade','Unemployment rate','Age at enrollment','Target']].corr()
          n (grade)','Father's qualification','Admission grade','Unemployment rate','Age at enrollment',"Target"]].corr(),annot=True,cmap='
Out[264]: <AxesSubplot:>
```



Conclusion

Higher grades strongly correlate with graduation, indicating a key predictor of success.

Younger enrollment age correlates with higher graduation rates.

While factors like unemployment rate, inflation rate, and GDP show weak correlations, they still offer insights into students' academic experiences within broader economic contexts.

Gender shows a weak positive correlation, while scholarships have a negative correlation, suggesting areas for targeted support.

While parents' qualifications don't directly correlate with student success, they still influence outcomes, highlighting the importance of familial support.

Specific marital statuses and age may impact academic success, suggesting the need for tailored support for different demographic groups.

Remove outliers beyond valid ranges (6329 outliers in Marital status and 9335 in Previous qualification) to ensure better accuracy. Require removing outliers