COS6008 Introduction to Data Science

Assignment 1, 2023, Semester 1

Exploratory Data Analysis of Higher Education Institution's Student Data

Student Details:

- Name: Thi Thanh Thuy Tran
- Student ID: 103514782
- Email: 103514782@student.swin.edu.au
- Submission Date: Friday 21 April 2023
- TuteLab Class: Thursday 17:30

Introduction

In this report, I will demonstrate the findings of Task 2 and Task 3 of the assignment, which involve exploratory data analysis of Higher Education Institution's Dataset on student enrollment. The dataset includes information on students' personal and academic backgrounds, enrollment status, and performance. The investigation' major findings will clarify potential factors influencing student enrolment and academic success.

1. Data Acquisition & Preparation

In Task 1, I acquired and prepared the data from three different files: “data1.cvs”, “data2.cvs”, “data3.cvs”.

1.1 Load

To begin with, I will start by loading the necessary Python libraries including pandas, numpy and matplotlib.pyplot and then I will load three CSV files into Pandas DataFrames. Next, I compared the loaded data to the original files to see if each loaded data set was equivalent to the data included in the raw data files. The equality check returns 'True’ 'for all datasets, it means that the loaded data sets are equivalent to the data contained in the raw data files.

1.2 Merge

Then I merged the three datasets into a single DataFrame using the unique identifier "ID".

I first merge data1 and data2 on the student ID, resulting in a data frame merged_data containing all the students described by the attributes in both data1 and data2. Next, I concatenate merged_data and data3, resulting in a data frame final_data containing all the students' records with a total of 38 attributes.

During data cleaning, I detected several data issues, including missing values, duplicates, impossible values, and extra whitespaces. These issues are handled by dropping duplicates, removing unnecessary columns and extra whitespaces, and filling in missing values with appropriate methods. For example, if a column or row has more than 50% missing values, I eliminated the entire column or row. In columns with fewer than 50% missing values, I replaced missing values with the mean of the non-missing values and filled missing values in categorical columns with the mode of the non-missing values. Moreover, when I was exploring data and created a histogram to see the distribution of ages at enrollment, I recognized a problem as some students whose age is greater than 100 which is an impossible value. To deal with this, I came back to this step to remove the rows where students whose age is greater than 100.

Some issues were found in merging the data, notably when dealing with the multiple data types and formatting anomalies in the three files. To resolve this issue, I thoroughly examined the data files and used proper data cleaning methods. Overall, I was successful in preparing a clean and usable dataset for further investigation. The combined and cleaned dataset had 4402 rows and 38 columns.

Task 2: Data Exploration

2.1 Exploring Categorical and Numerical Columns

In this part, I first explore one categorical and one numerical column. For a categorical column, the "Marital status" column is chosen. This column contains information about students' marital status, which may affect their academic performance or financial situation. Figure 1 is a bar chart I created to visualize the "Marital status" column, which indicates the ount of students based on their marital status.

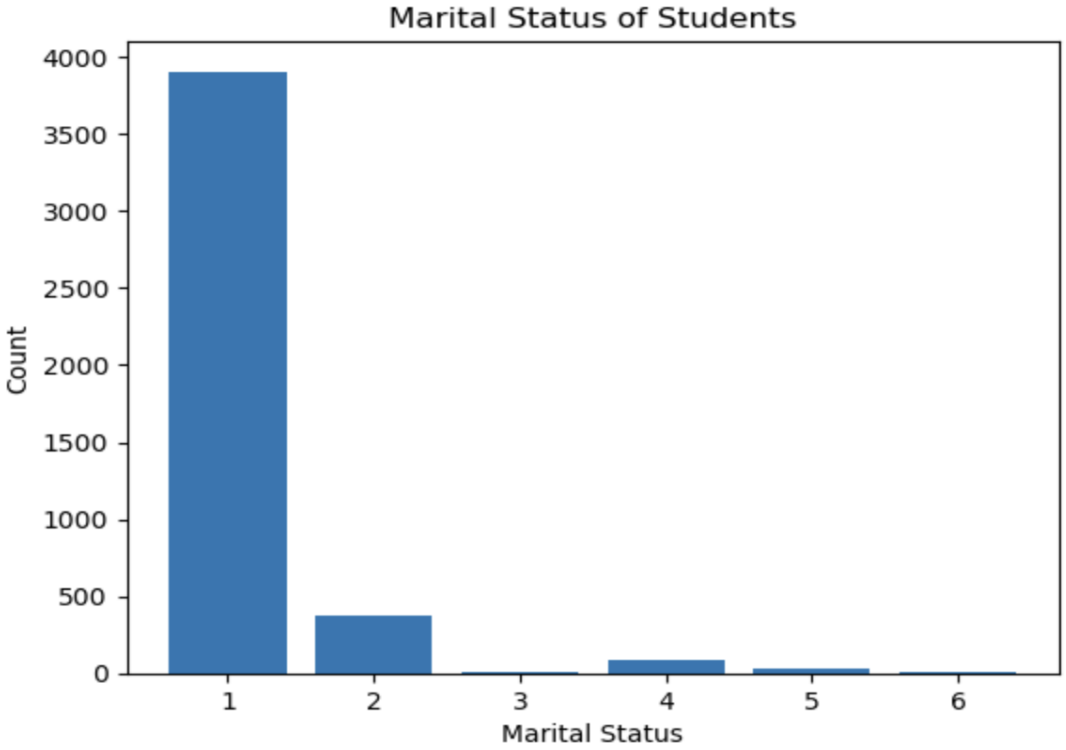


Figure 1: Marital Status of Students

The "Age at enrolment" column was chosen for a numerical column because it contains the age of students at the time of enrolment and can potentially be useful in analyzing student performance in relation to their age. Figure 2 shows the outcome of creating a histogram to see the distribution of ages at enrolment to visualize this column.

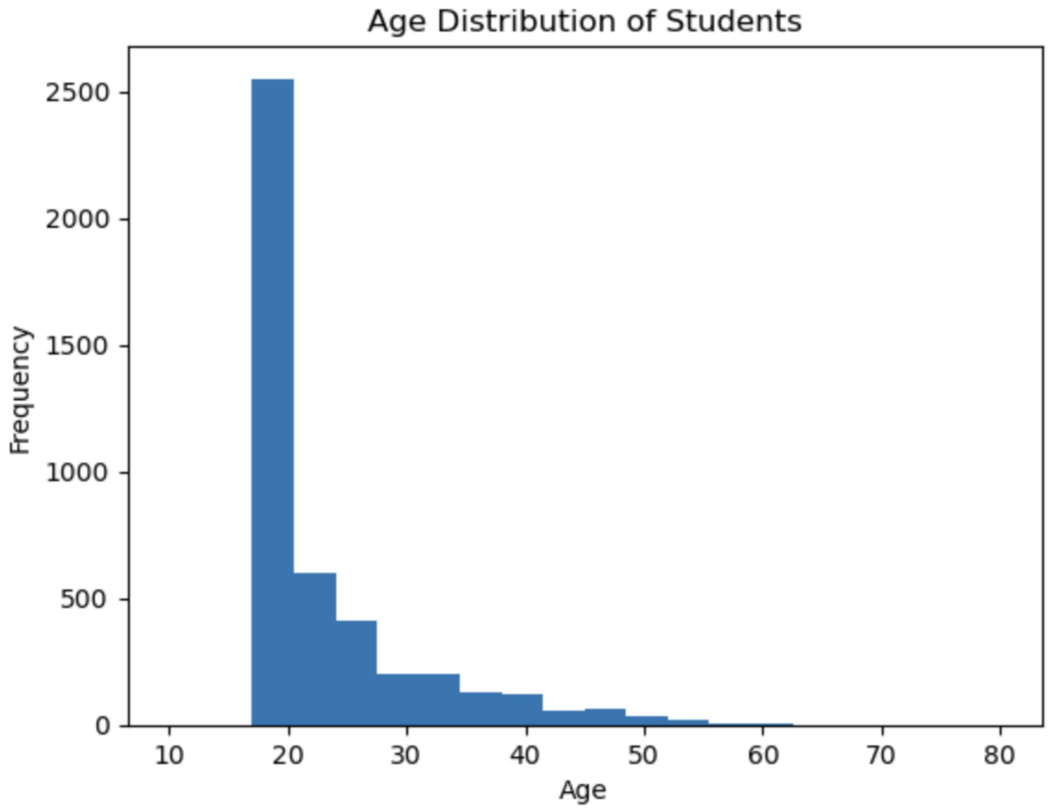


Figure 2: Age Distribution of Students

Overall, these plots aid in understanding the distribution of students based on marital status and age at enrolment, which can provide insight into academic achievement and other factors influencing their education.

2.2 Exploring Column Relationships

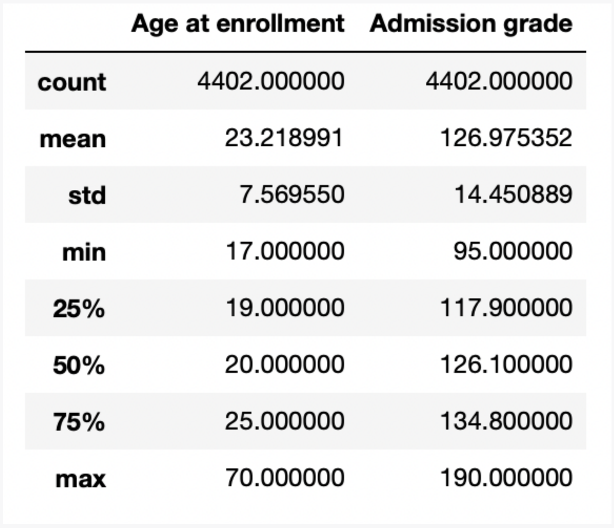
Next, in order to explore the relationship within pairs of columns, I had made a hypothesis about their possible relationship. Here are three pairs that I was planning to explore:

1. Age at Enrollment and Admission Grade
2. Educational special needs and admission grade
3. Enrolled Units and Grade

I chose these pairs to explore potential relationships that could be useful in predicting student performance. For each pair, I computed the descriptive statistics to get an overview of the distribution of each column and created visualization tools to see their relationship.

2.2.1 Age at Enrollment and Admission Grade

For the first pair, I hypothesized a negative relationship between enrolment age and admission grade; as enrolment age grows, admission grade is likely to fall. First, let's have an overview at the descriptive statistics in figure 3, which show that the average age at enrolment is 23.3 years and the average admission grade is 126.09. The standard deviation for entrance grade is about double that of age at enrolment, indicating that the data is more spread out. The lowest age is 17 years old, and the maximum age is 70 years old, which is quite old.



	Age at enrollment	Admission grade
count	4402.000000	4402.000000
mean	23.218991	126.975352
std	7.569550	14.450889
min	17.000000	95.000000
25%	19.000000	117.900000
50%	20.000000	126.100000
75%	25.000000	134.800000
max	70.000000	190.000000

.

Figure 3: Descriptive statistics of Age at enrollment and Admission grade

Next a scatter plot was used to visualize the relationship between the two columns and calculated the correlation coefficient. My findings in figure 4 revealed that there is a weak negative correlation between the age at enrollment and admission grade, with the correlation coefficient being -0.03049.

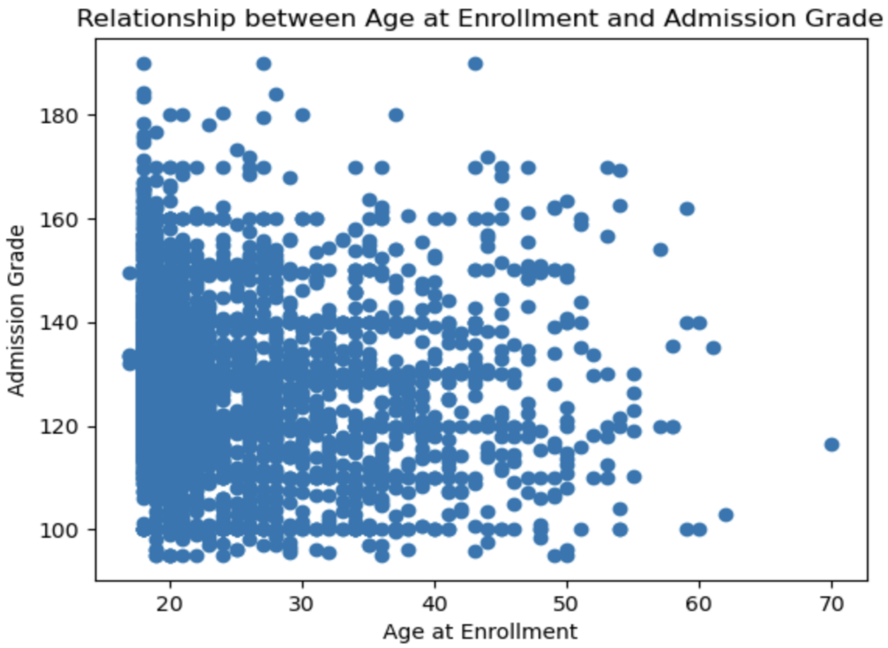


Figure 4: Relationship between Age at Enrollment and Admission Grade

2.2.2 Educational special needs and admission grade

I hypothesized that students with special needs would have lower admission grades. To explore the relationship, I created a box plot as shown in figure 5 for these columns as it can help to compare the distribution of these variables for students with and without educational special needs.

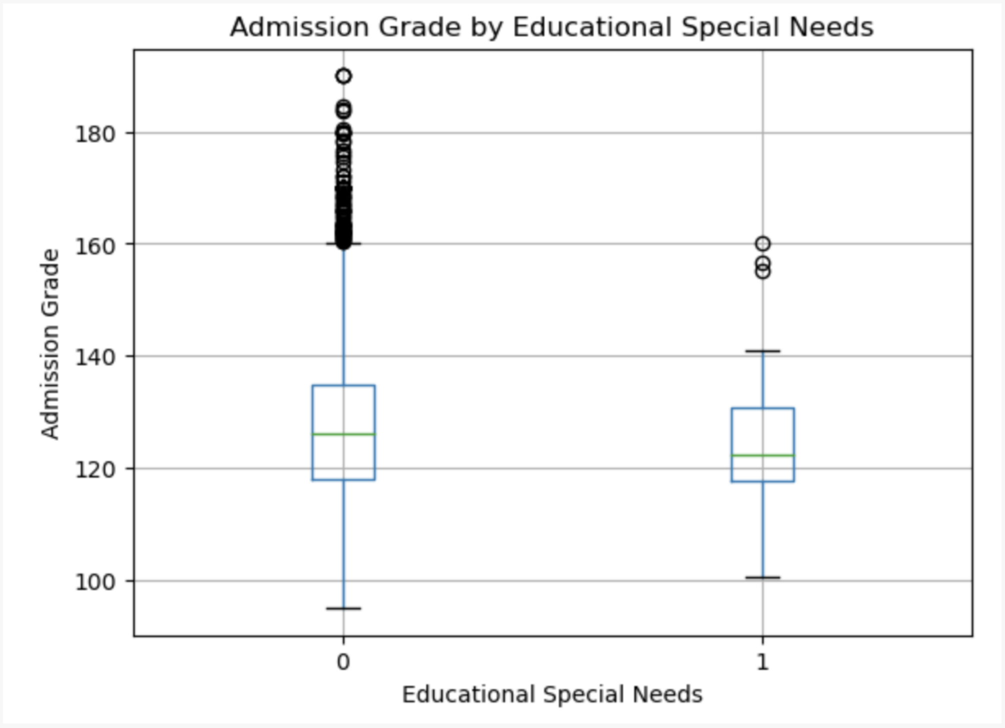
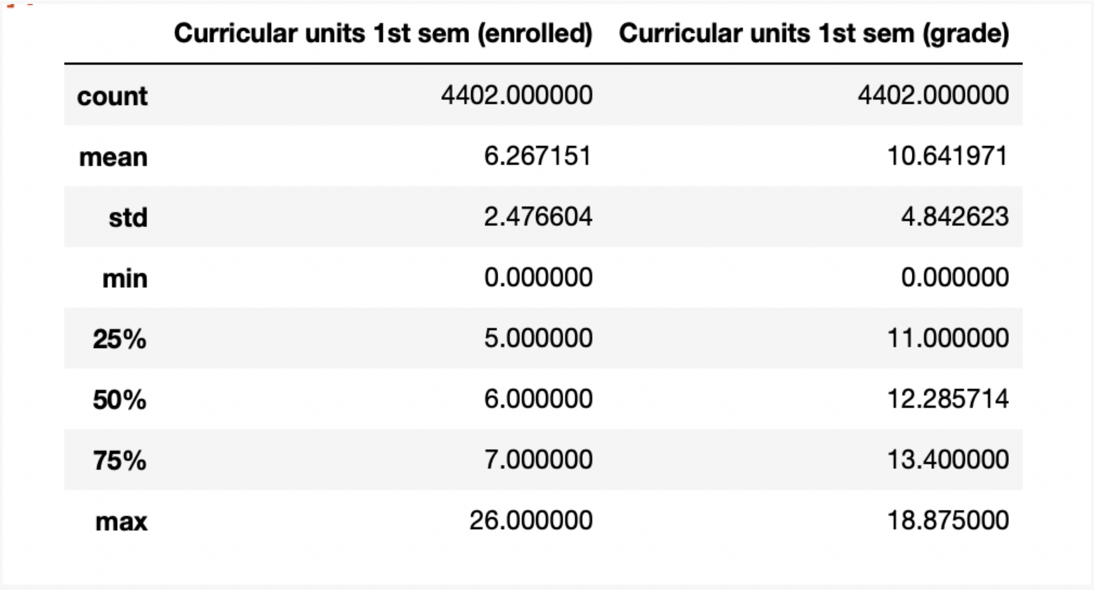


Figure 5: Admission Grade by Educational Special Needs

According to the box plot, the median admission grade for students with educational special needs is slightly lower than for students without special needs. We can also notice that children with exceptional needs have a larger dispersion than students without special needs. This indicates that there is more variability in the admission grades for students with special needs.

2.2.3 Enrolled Units and Grade

I hypothesized that there might be a positive association between the variables Enrolled Units and Grade. Descriptive statistics in figure 6 demonstrates the average number of units enrolled is 6.27, and the average grade is 10.64. This implies that the average student enrolls in approximately 6 units during their first semester.



	Curricular units 1st sem (enrolled)	Curricular units 1st sem (grade)
count	4402.000000	4402.000000
mean	6.267151	10.641971
std	2.476604	4.842623
min	0.000000	0.000000
25%	5.000000	11.000000
50%	6.000000	12.285714
75%	7.000000	13.400000
max	26.000000	18.875000

Figure 6: Descriptive statistics between Enrolled Units and Grade in 1st sem

According to figure 7, the scatter plot revealed a weak association, but the correlation value of 0.39 indicates that these two variables have a moderately positive link. The average grade tends to rise as the number of enrolled units rises.

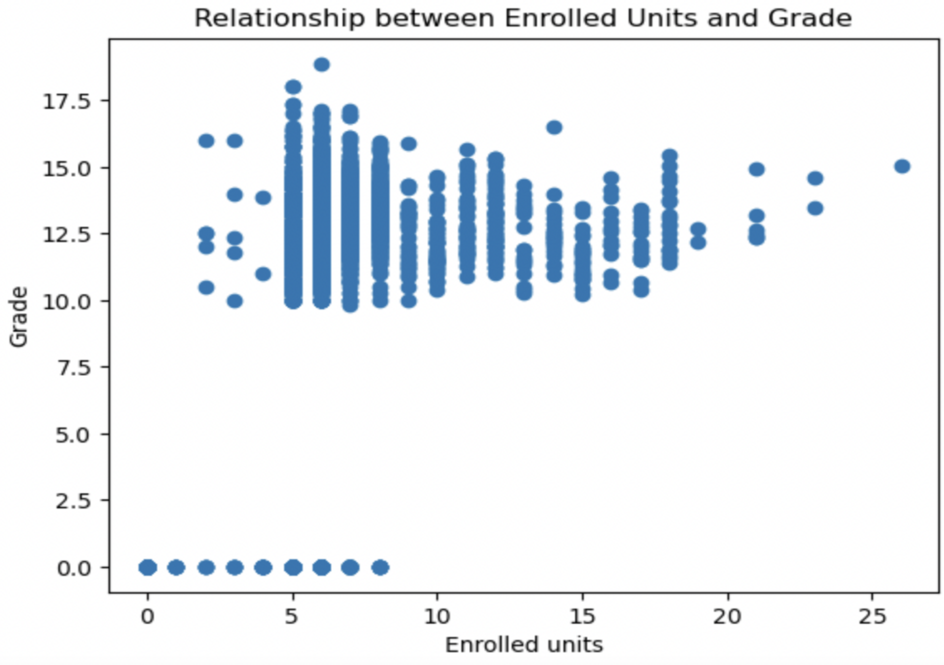


Figure 7: Relationship between Enrolled Units and Grade

2.3 Exploring Multiple Numerical Columns

In this task,I built a scatter matrix with diagonal histograms and scatter plots for each pairwise combination of the 6 columns chosen. The scatter matrix is used to visualize the relationship between the numerical variables and identify potential correlations or patterns. Six columns are chosen explore if there were any significant relationships between the different academic measures are:

1. Admission grade: This column is critical for determining students' academic performance at the time of admission.
2. Age at enrollment: This column may shed light on any association that exists between age and academic performance.
3. Curricular units 1st sem (credited): This column indicates the number of curricular units that the student was credited for in their first semester which may affect student academic performance.
4. Curricular units 2nd sem (credited): This column represents the number of curricular units that the student was credited for in their second semester which may affect student academic performance.
5. Unemployment rate: This column indicates the country's unemployment rate, which may have an impact on students' financial and academic success.
6. Inflation rate: This column indicates the country's inflation rate, which may have an impact on students' financial and academic success.

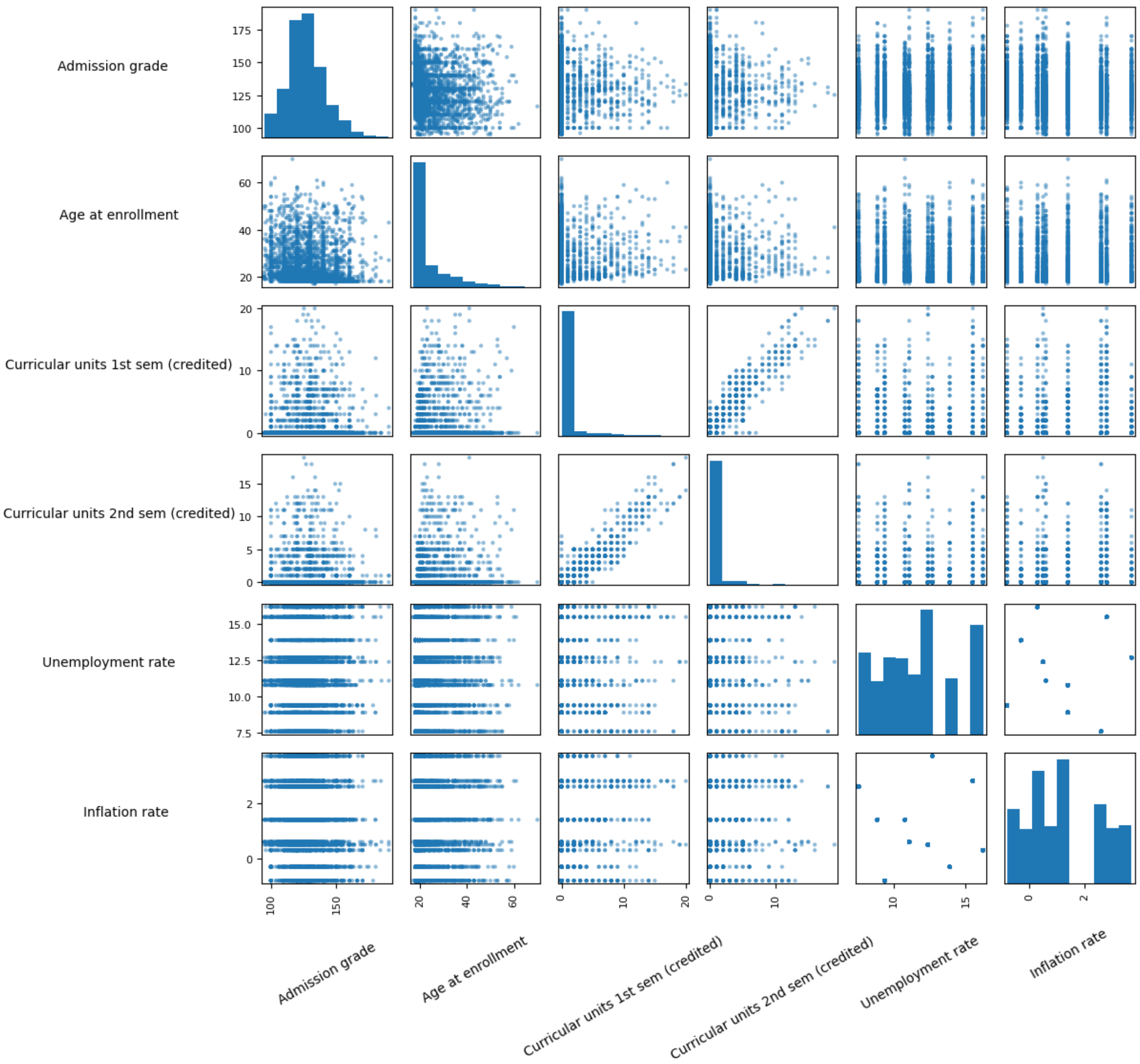


Figure 8: Scatter matrix with six numerical columns

Figure 8 demonstrates a visual representation of the relationships between the selected numerical columns. For example, we can see that there is a positive correlation between the Curricular units 1st sem (credited) and the Curricular units 2nd sem (credited) indicating the more units that the student was credited for 1st sem, the more likely they get for 2nd sem, while there is no clear relationship for other variables. To indicate further relationships between other variables, we might need to use more visualization tools. Additionally, by using scatter matrix, we can visually identify any outliers or unusual patterns in the data.

Conclusion

To sum up, this report has demonstrated the findings of the data exploration and analysis of the student performance dataset. The report accomplished the responsibilities of data gathering, cleaning, and merging, as well as data exploration via various visualizations and descriptive statistics. The analysis identified several interesting patterns and relationships between the variables, providing insights into the factors that influence student performance. Overall, the report emphasizes the significance of data exploration and analysis in generating important insights into large datasets.