Comprehensive Data Exploration Report

1. Introduction

This report presents the findings from a comprehensive data exploration aimed at predicting student dropout rates. The dataset contains information on students' socio-economic background, academic performance, and institutional factors that may influence dropout decisions.

2. Data Overview

The dataset consists of 4424 entries and 37 columns covering various attributes such as marital status, application mode, previous qualification, grades, socio-economic factors (e.g., unemployment rate), and the target outcome (dropout, graduate, or enrolled). Key features include the admission grade, age at enrollment, and curricular units across two semesters.

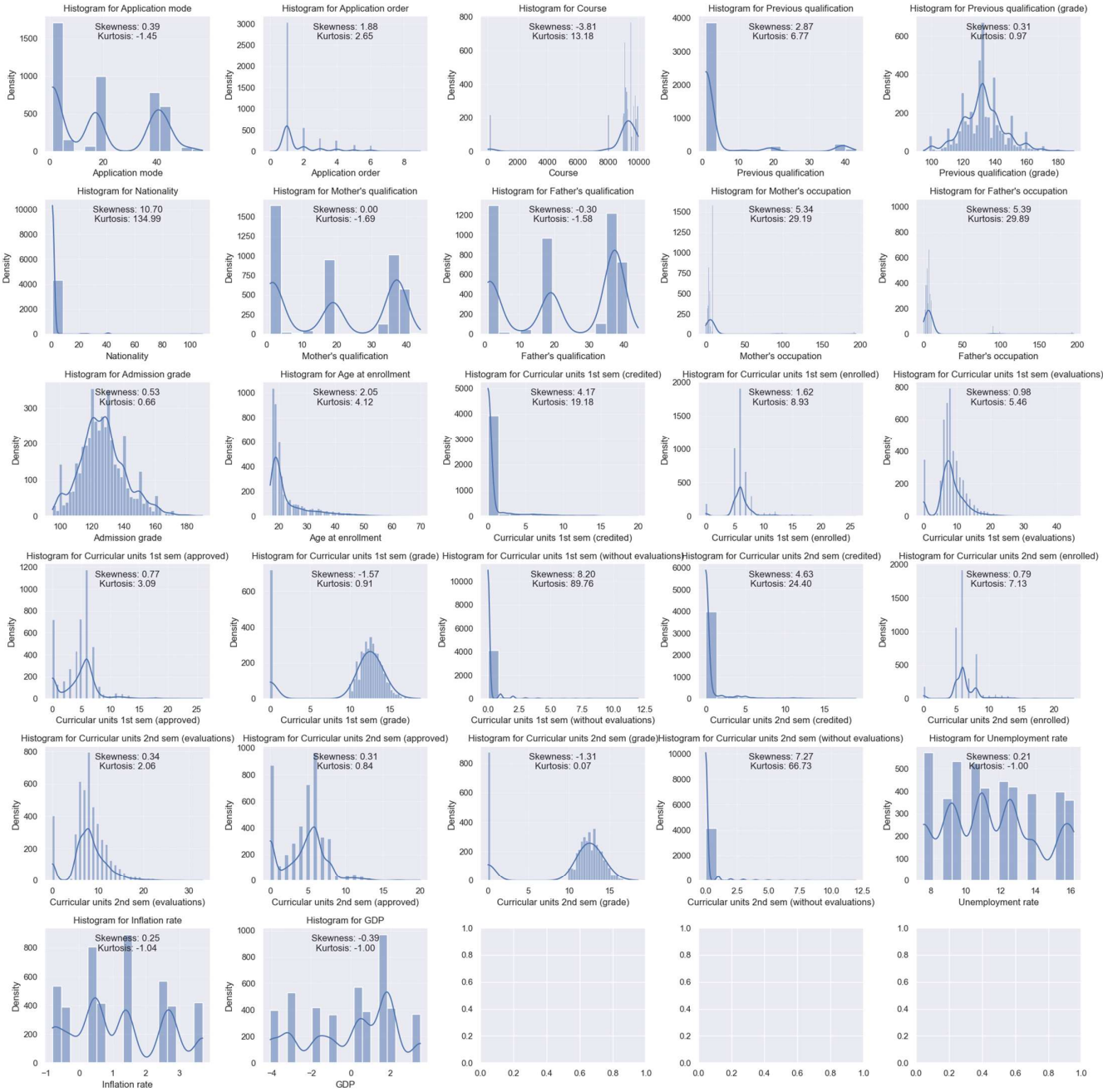
3. Data Preprocessing

To prepare the dataset for analysis, categorical columns were converted into appropriate data types. Outliers in numerical columns were identified.

4. Exploratory Data Analysis

4.1 Univariate Analysis

Various features were explored individually to understand their distributions. Histograms and boxplots were generated for numerical features, and countplots for categorical features.



Histogram of Variables in the dataset

From the histogram, the following can be deduced;

The dataset exhibits various patterns of skewness across different features, offering insights into student demographics and external factors. Several variables, such as application mode and admission grade, show slight right skewness, indicating fairly even distributions with peaks in

specific ranges. In contrast, features like application order, age at enrollment, and previous qualifications display strong right skewness, suggesting most students fall into lower ranges, with fewer outliers at higher values. This is particularly evident in age and curricular units, where younger students and those with fewer course credits dominate the dataset.

Interestingly, course selection and inflation rate have significant left skewness, implying that students cluster around certain courses, with inflation rates showing lower values for most periods. Parental qualifications, however, are more symmetrically distributed, indicating a balanced spread in educational backgrounds. Economic indicators, such as GDP and unemployment rate, reflect a slightly right-skewed distribution, with lower unemployment rates and balanced GDP levels, though some outliers exist.

Overall, these patterns suggest that student characteristics such as age, application behaviors, and economic factors like inflation and unemployment may play a role in predicting student outcomes, particularly in dropout risk.

After the histogram and boxplot for all univariate variables was plotted, it was found out that most of the columns were categorical. These columns was transformed to the appropriate categorical type.

Countplot of all categorical variables in the dataset

From the countplot above, it can be deduced that;

Most students are single, indicating that the majority are likely younger and unmarried. In terms of application mode, while there are multiple options, one stands out as the most popular, suggesting a preferred way of applying. Similarly, a large portion of students apply during the first application order, hinting at either the benefits of applying early or a general preference for doing so.

When it comes to academic engagement, most students exhibit full attendance, which could be tied to better academic performance or a lower risk of dropping out. A significant finding is that many students' parents have low or no formal qualifications, which might reflect socio-economic barriers that influence their children’s academic success.

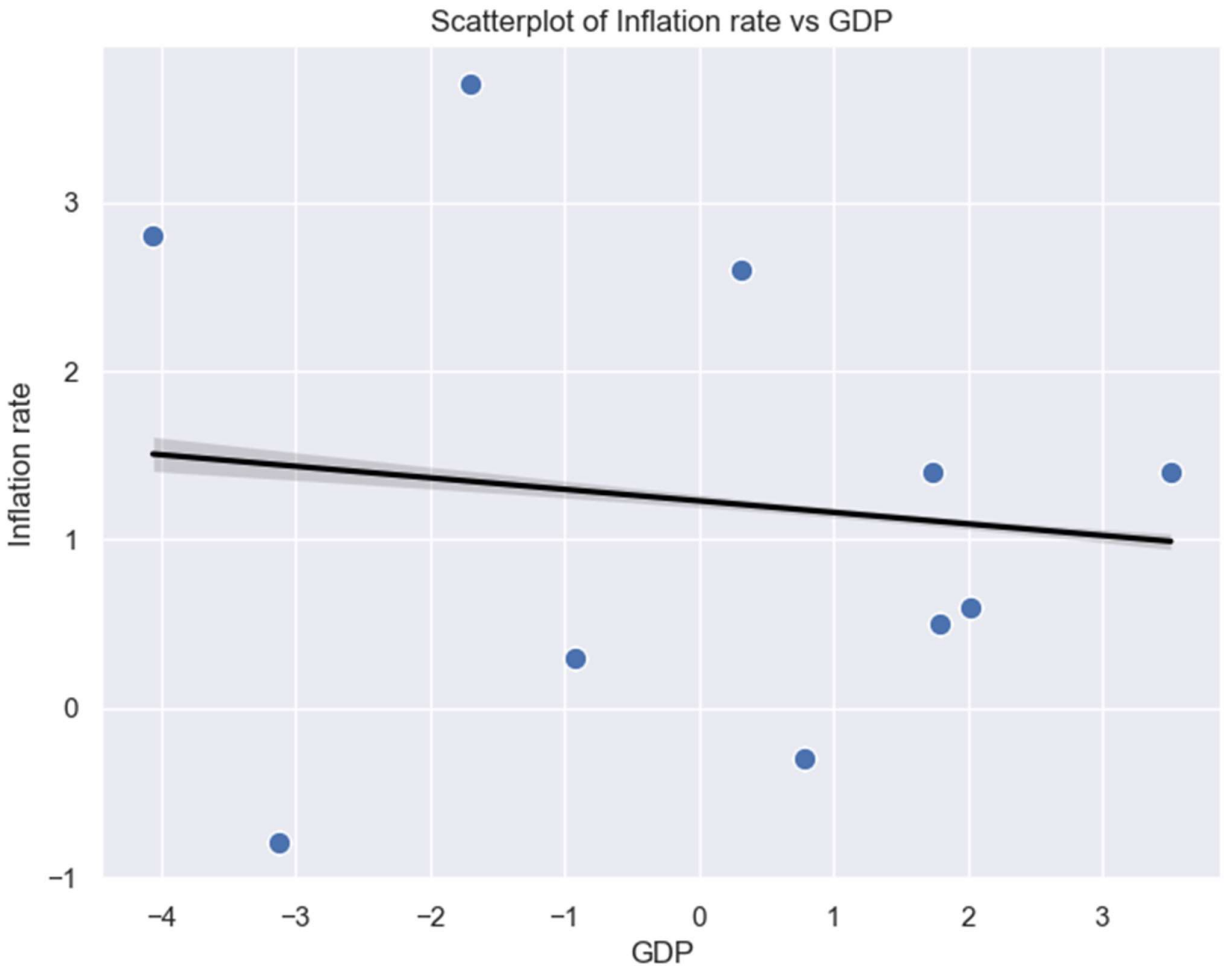
In terms of special educational needs, only a small minority of students require such support, meaning most of the population doesn’t face those additional challenges. However, a considerable number of students have unpaid tuition fees, which might serve as a strong predictor of dropout, as financial instability often plays a role in student retention. Similarly, while many students have paid their fees on time, a notable portion has not, reinforcing the potential link between financial stability and staying in school.

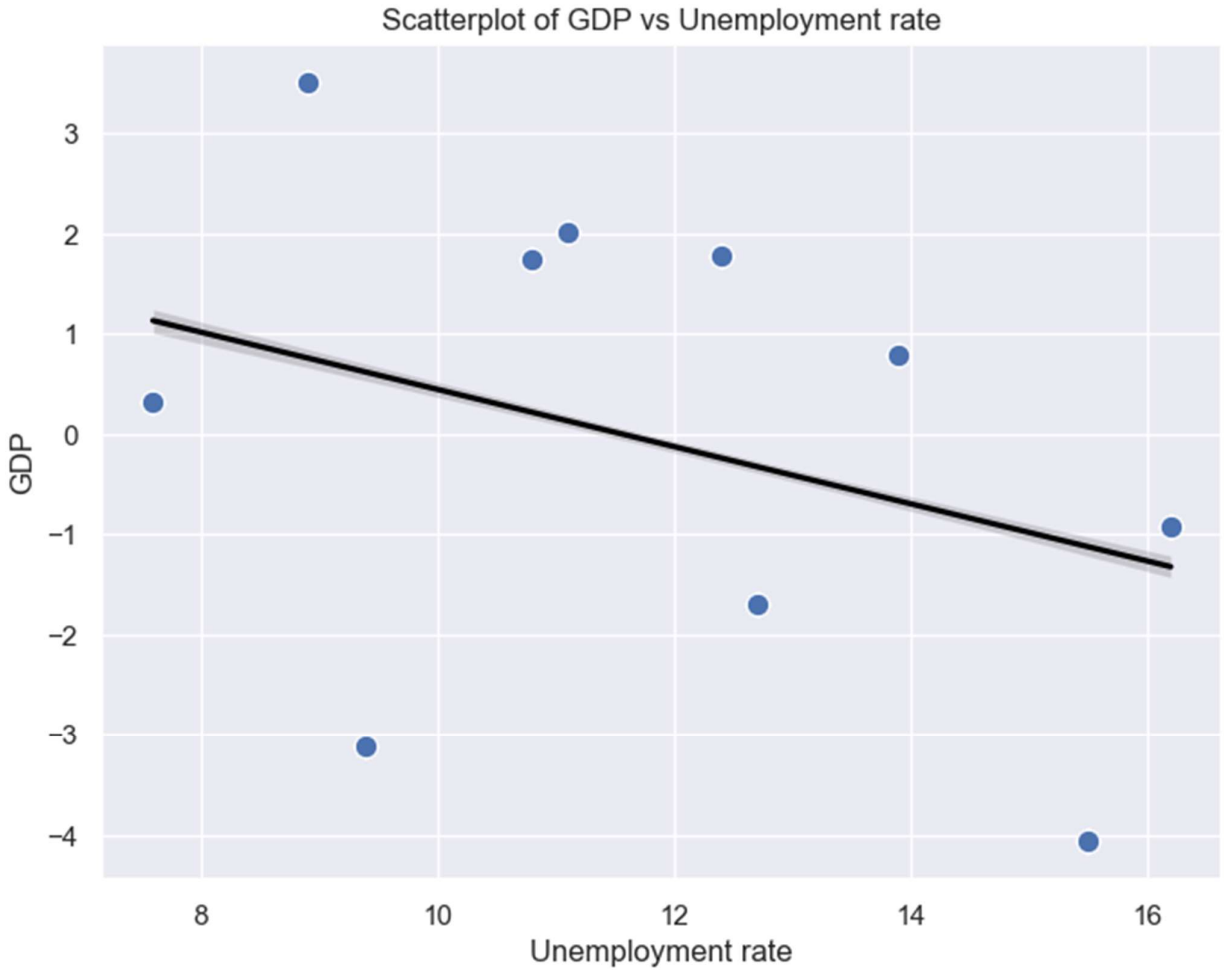
A significant number of students are not on scholarships, which may contribute to their financial challenges, and although most students eventually graduate, the number of dropouts is still substantial, signaling an ongoing issue. Financial struggles, parental education, and application

timing all appear to play significant roles in influencing whether students stay in school or drop out.

4.2 Bivariate Analysis

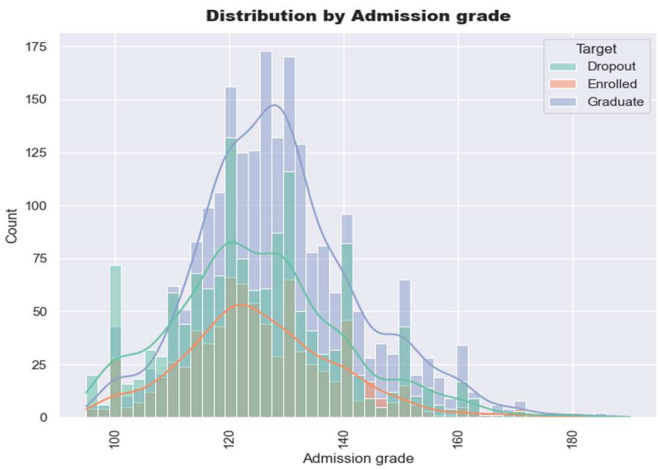
Scatterplots and regression analysis were used to explore relationships between pairs of variables. Notable findings include a weak negative correlation between GDP and Inflation rate, and a negative correlation between unemployment rate and GDP.

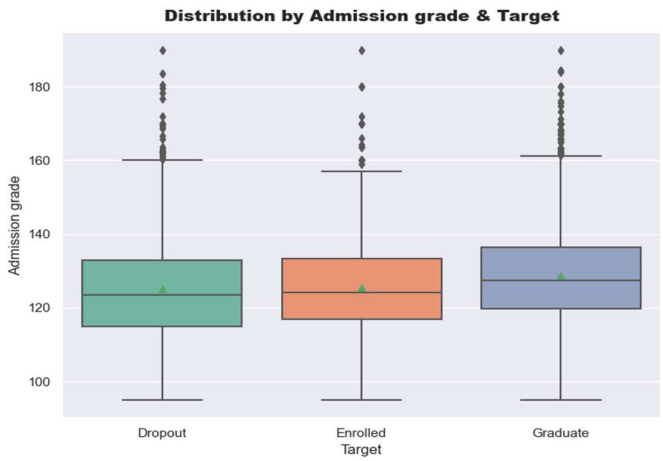




Plots were also created for numerical variables against categorical values, they include

Admission grade vs Target:



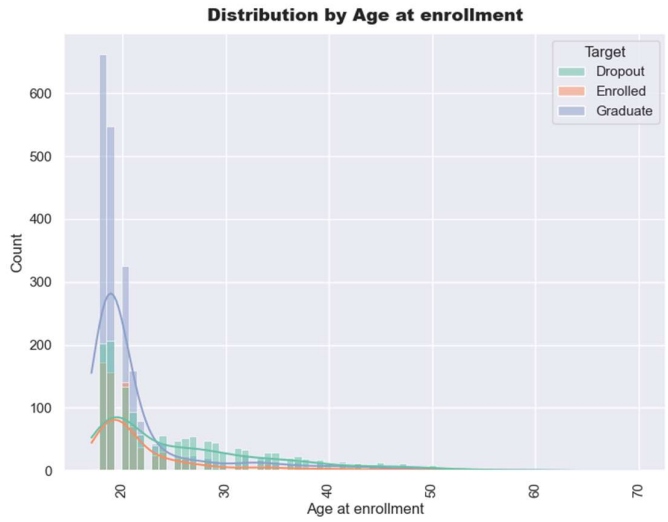


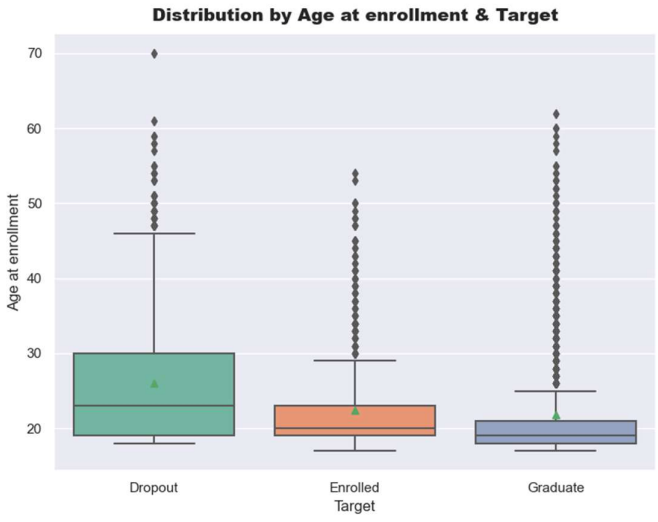
Summary:

- Graduates typically have higher admission grades, clustering around 120-160, with the highest densities in the 130-150 range.
- Dropouts have a wider spread, with many having lower grades around 100-140.
- Enrolled students show a more balanced, lower distribution, with fewer peaks compared to graduates or dropouts.
- Graduates have the highest median admission grade, followed by enrolled students, with dropouts having the lowest.

Higher admission grades seem to lead to better outcomes, with graduates consistently having higher scores. Students with lower grades are more likely to drop out.

Age of Enrollment vs Target





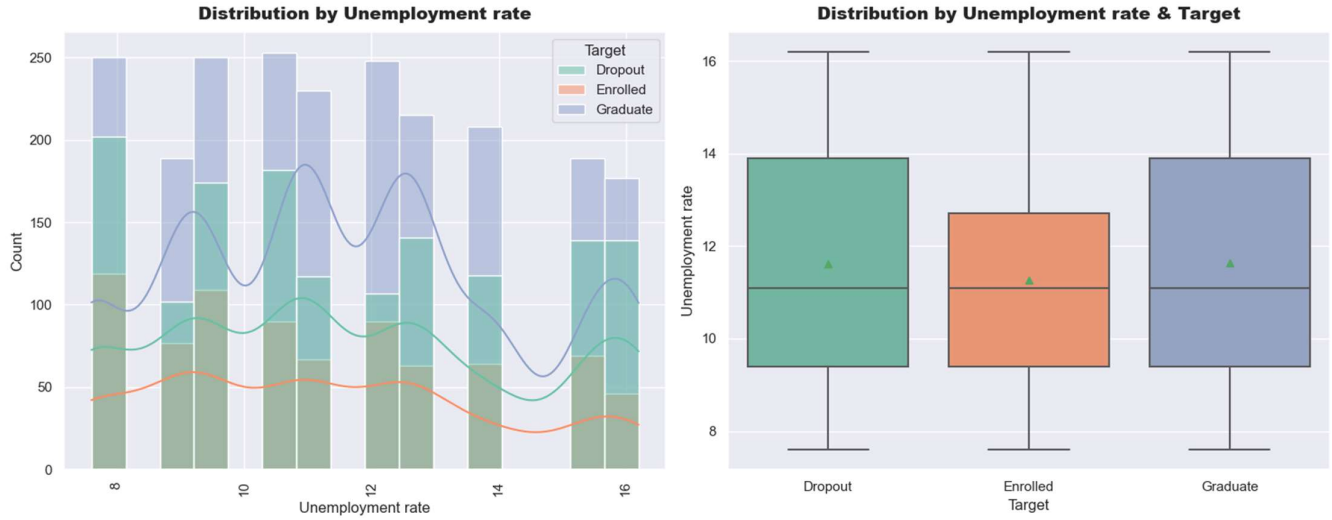
Summary:

- Most students, regardless of outcome, enroll in their early 20s, with a sharp decline in enrollment as age increases.

- Graduates dominate the younger age groups, especially around 18-22. typically enroll in their early 20s, similar to enrolled students, but there are outliers up to age 60.
- Dropouts are more common among older age groups compared to graduates.They also tend to enroll later, with a median age around 25 and some enrolling even into their 60s.
- Enrolled students have a fairly even, though smaller, presence across all age ranges.They are younger, with a median age of around 21.

Younger students are more likely to graduate, while older students show higher dropout rates.

Unemployment Rate vs Target



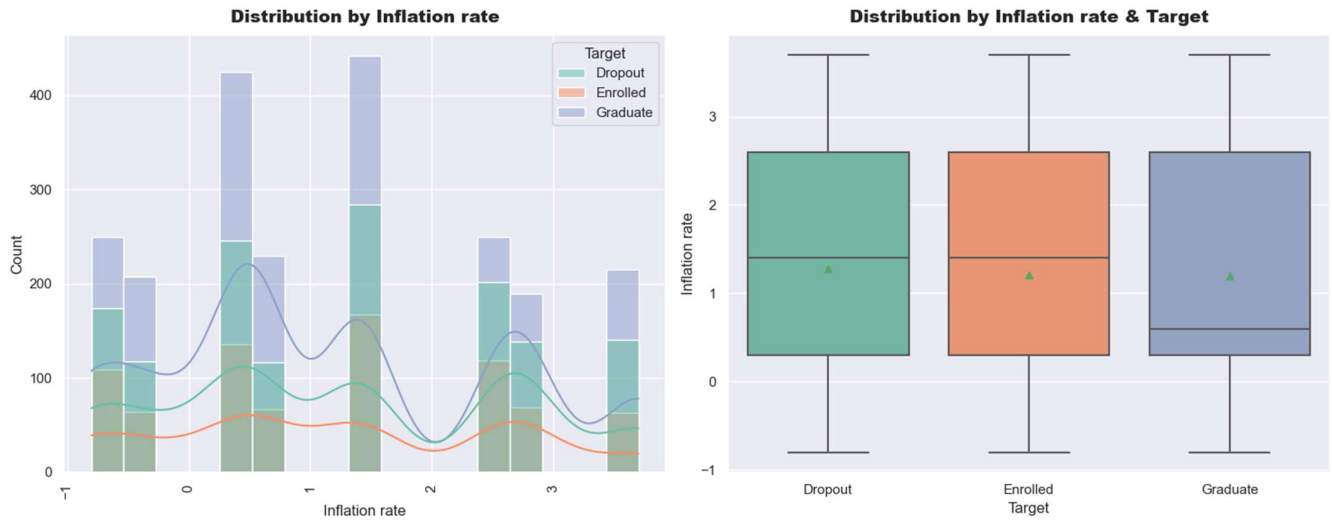
Summary:

- Graduates tend to come from areas with higher unemployment (10-16%), possibly because tough job markets keep students in school longer.
- Dropouts are spread across different unemployment levels, with a slight concentration around mid-level rates (10-14%).

- Enrolled students consistently experience around 11-12% unemployment, showing less variation than the other groups.

In summary, higher unemployment may push students to graduate, while dropouts happen across a range of unemployment conditions.

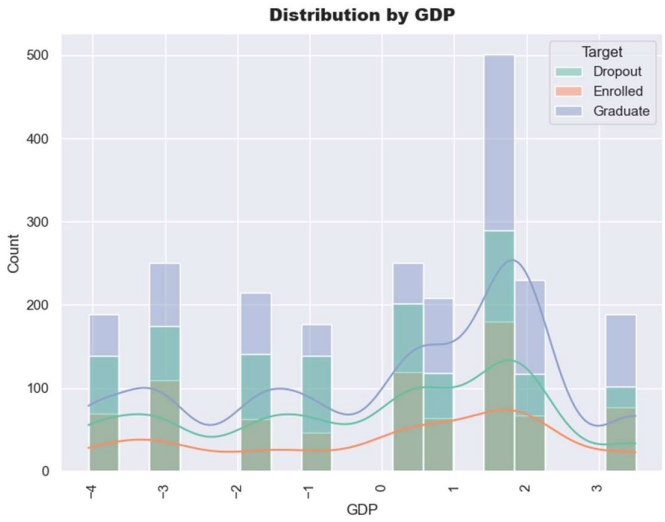
Inflation rate vs Target

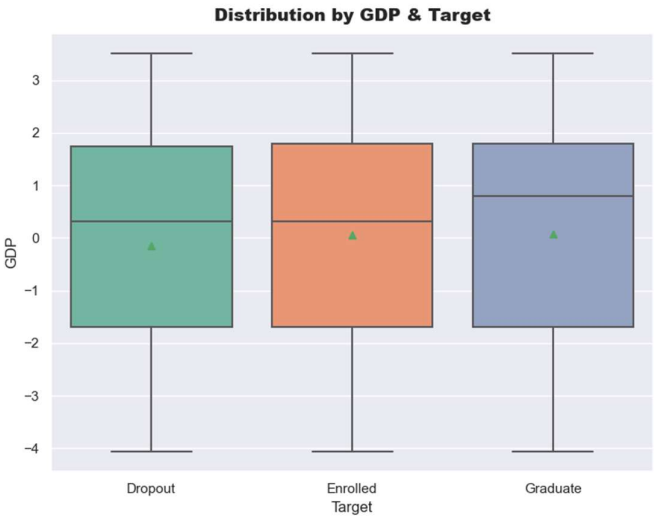


Summary:

- Dropouts tends to have a slightly lower average inflation rate compared to the other groups.
- Enrolled students distribution is somewhat similar to Dropout, but with a slightly higher average inflation rate.
- Graduates seems to have the highest average inflation rate, suggesting that individuals or entities experiencing higher inflation rates are more likely to be graduates.

GDP vs Target

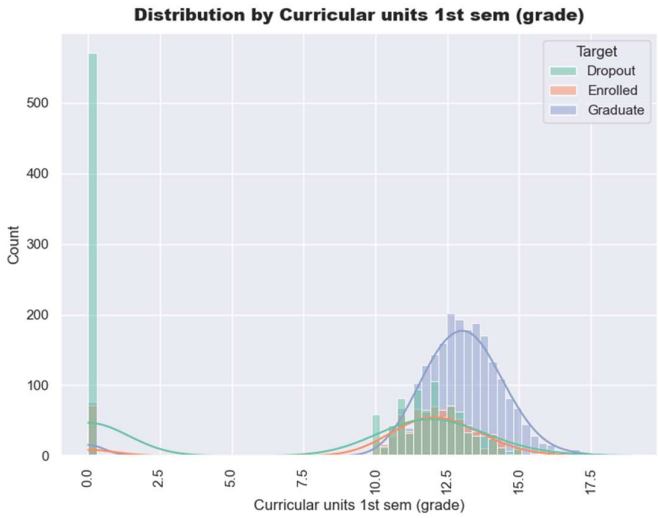


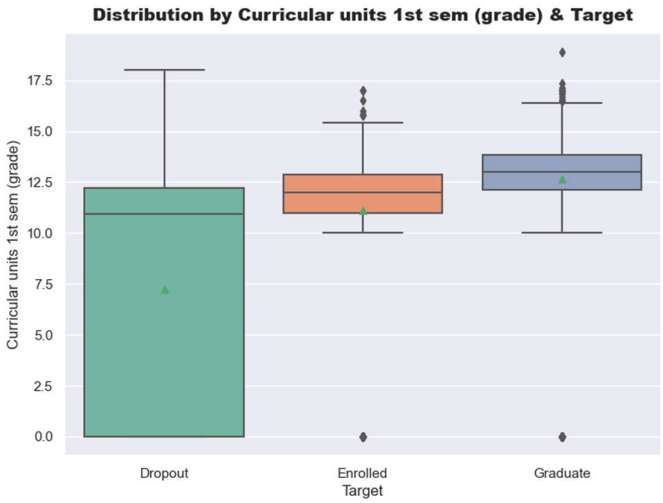


Summary:

- Dropouts appears to have a slightly lower average GDP compared to the other groups.
- Enrolled sudent's distribution is somewhat similar to Dropout, but with a slightly higher average GDP.
- Graduate seems to have the highest average GDP, suggesting that individuals or entities with higher GDP are more likely to fall into this target category. The mean GDP on average increases from dropout to Graduate with Graduates having the higher mean GDP which makes the GDP a good predictor of the 'Target' variable.

Curricular units 1st sem vs Target



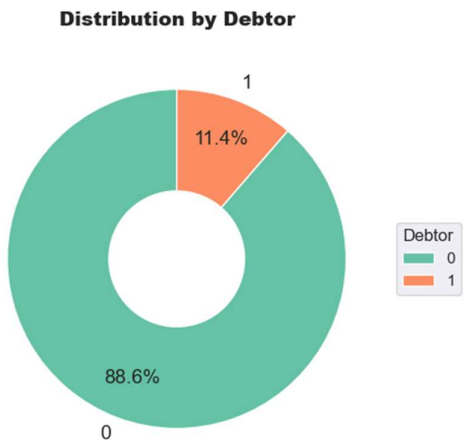


Summary:

- Dropouts have a slightly lower average grade compared to the other groups.
- Enrolled students distribution is somewhat similar to Dropout, but with a slightly higher average grade.
- Graduates seems to have the highest average grade, suggesting that students with higher grades are more likely to fall into this target category.

Dropouts on average, tends to have a lower 1st semester grade than the other categories.

Debtor Status vs Target

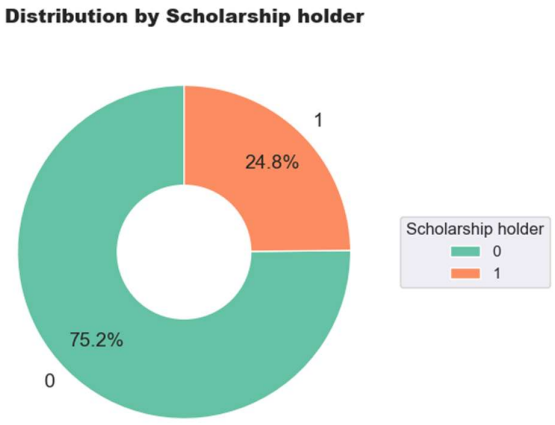


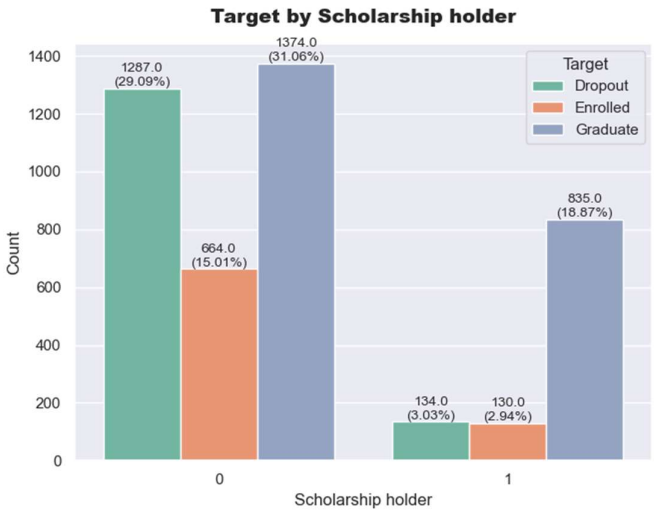
Summary:

The majority of students are not debtors i.e 88.6% of the students while the remaining 11.4% are debtors. For the category of students who are not debtors, the graduates are the most frequent followed by the dropout category of students. For the category of students who owe, this observation is reversed with the dropout students being the most frequent.

This shows that majority of students who are on not debtors tends to graduate.

Scholarship Holder vs Target





Summary:

The majority of students are not scholarship holders i.e 75.2% of the students while the remaining 24.8% are scholarship holders. For the category of students who are not scholarship holders, the graduates (31.06%) are the most frequent closely followed by the dropout category of students (29.09%). For the category of students who are scholarship holders, the graduate category still remains the most frequent.

This shows that majority of students who are on scholarships tends to graduate.

After this set of visualizations was carried out, a new variable was created called drop_stats to hold the value of 1 if the student was a dropout and 0 if otherwise. This ensured that the hypothesis testing went on smoothly. Another variable called the 'Average curricular units which signified the average of the first and second semester units grades.

5. Key Insights and Findings

6. Hypothesis Testing

Test Name	**Test Type**	**Hypothesis**	**Test Statistic**	**p-value**	**Conclusion**
Unemployment Rate vs Dropout Status	Correlation Test	H_0: r = 0 H_1: r ≠ 0	0.01298	0.388079	Fail to reject H0

Gross Domestic Product vs Dropout Status	Correlation Test	H_0: r = 0 H_1: r ≠ 0	0.04623	0.002059	Reject H0
Inflation Rate vs Dropout Status	Correlation Test	H_0: r = 0 H_1: r ≠ 0	0.06422	0.064223	Fail to reject H0
Test of Association between Debt status and Dropout status	Chi-Square Test	H_0: The variables are independent H_1: The variables are not independent	232.825	3.25E-49	Reject H0
Test of Association Between Tuition fees status and Dropout status	Chi-Square Test	H_0: The variables are independent H_1: The variables are not independent	814.764	4.9E-175	Reject H0
Admission grade for dropouts vs non dropouts	T-Test	H0: $\mu_1 = \mu_2$ H1: $\mu_1 < \mu_2$	-6.237	2.59E-10	Reject H0
Test of Association between Scholarship holder and Dropout rate	Chi-Square Test	H_0: The variables are independent H_1: The variables are not independent	266.318	1.98E-56	Reject H0

Test of Association between International and Dropout rate	Chi-Square Test	H_0: The variables are independent H_1: The variables are not independent	0.4748	0.956	Fail to reject H0
Test of Association between Educational Special Needs and Dropout rate	Chi-Square Test	H_0: The variables are independent H_1: The variables are not independent	0.0348	0.999	Fail to reject H0
Average Curricular Units for dropouts vs non dropouts	T-Test	H0: $\mu_1 = \mu_2$ H1: $\mu_1 < \mu_2$	-444.63	0.000	Reject H0

Multiple hypotheses were tested using Chi-square and T-tests. Notable results include:

i. **Unemployment Rate vs Dropout Status:**
There is no evidence to suggest a significant relationship between the unemployment rate and the likelihood of dropping out. Therefore, we fail to reject the null hypothesis that the correlation is zero, implying that unemployment rate does not seem to influence student dropout rates based on this data.

ii. **GDP vs Dropout Status:**
There is evidence to suggest a weak but statistically significant correlation between GDP and dropout rates. Based on the data, we reject the null hypothesis (H0), implying that GDP does have a small effect on the likelihood of students dropping out.

iii. **Inflation Rate vs Dropout Status:**
There is evidence to suggest a weak but statistically significant correlation between GDP and dropout rates. Based on the data, we reject the null hypothesis (H0), implying that GDP does have a small effect on the likelihood of students dropping out.

iv. **Debt Status vs Dropout Status:**

This means that there is a strong association between debt status and dropout status. Phi's coefficient measures the strength of the association between the two variables. A value of 0.22941 indicates a moderate positive association. This means that as debt status increases, the likelihood of dropping out also tends to increase.

v. **Tuition fees payment status vs Dropout Rate:**

Given the p-value is much smaller than 0.05, we reject the null hypothesis (H0). This implies that there is a statistically significant relationship between "Tuition fees up to date" status and "Dropout status" — the two variables are not independent.

The Phi's Coefficient of -0.42915 indicates a moderate negative association between the two variables. The negative sign suggests that individuals with up-to-date tuition fees are less likely to drop out, while those with unpaid fees are more likely to drop out.

vi. **Admission grade of Dropouts vs Admission grade of Non Dropouts:**

The p-value (2.59e-10) is extremely small and well below the significance level ($\alpha = 0.05$), indicating that the difference is statistically significant.

The evidence strongly supports the alternative hypothesis. We reject the null hypothesis (H0) and conclude that students who dropped out had significantly lower admission grades compared to those who did not.

vii. **Scholarship Holder vs Dropout Status:**

The extremely small p-value (1.983E-56 is far below the significance level ($\alpha = 0.05$), indicating a highly significant result.

Phi's coefficient is a measure of association strength for Chi-square tests. A value of -0.24535 suggests a moderate negative association between the two variables (e.g., as scholarship rate increases, the likelihood of dropout may decrease).

viii. **International vs Dropout Status**

The high p-value indicates that we fail to reject the null hypothesis. International student status does not appear to have a significant effect on dropout rates. International and domestic students are likely to have similar dropout tendencies.

ix. **Educational Special Needs vs Dropout status**

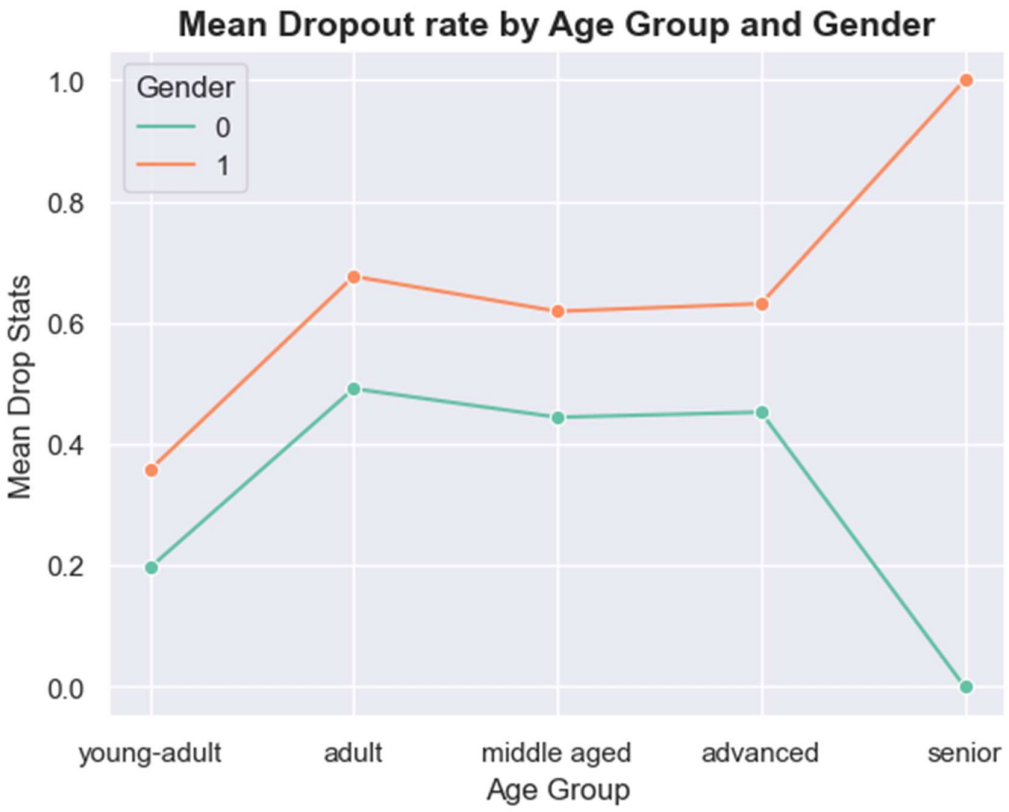
The p-value is very high, meaning we fail to reject the null hypothesis. This suggests that having special educational needs does not significantly affect dropout rates. Students with or without special educational needs show similar patterns in terms of dropout behavior.

x. **Average curricular units grade of Dropouts vs Admission grade of Non Dropouts**

The p-value is extremely small, leading us to reject the null hypothesis. Dropouts tend to complete significantly fewer curricular units compared to non-dropouts, reinforcing the idea that academic progress is strongly linked to dropout behavior.

7. Multivariate Analysis

Mean Dropout rate by Age group and Gender



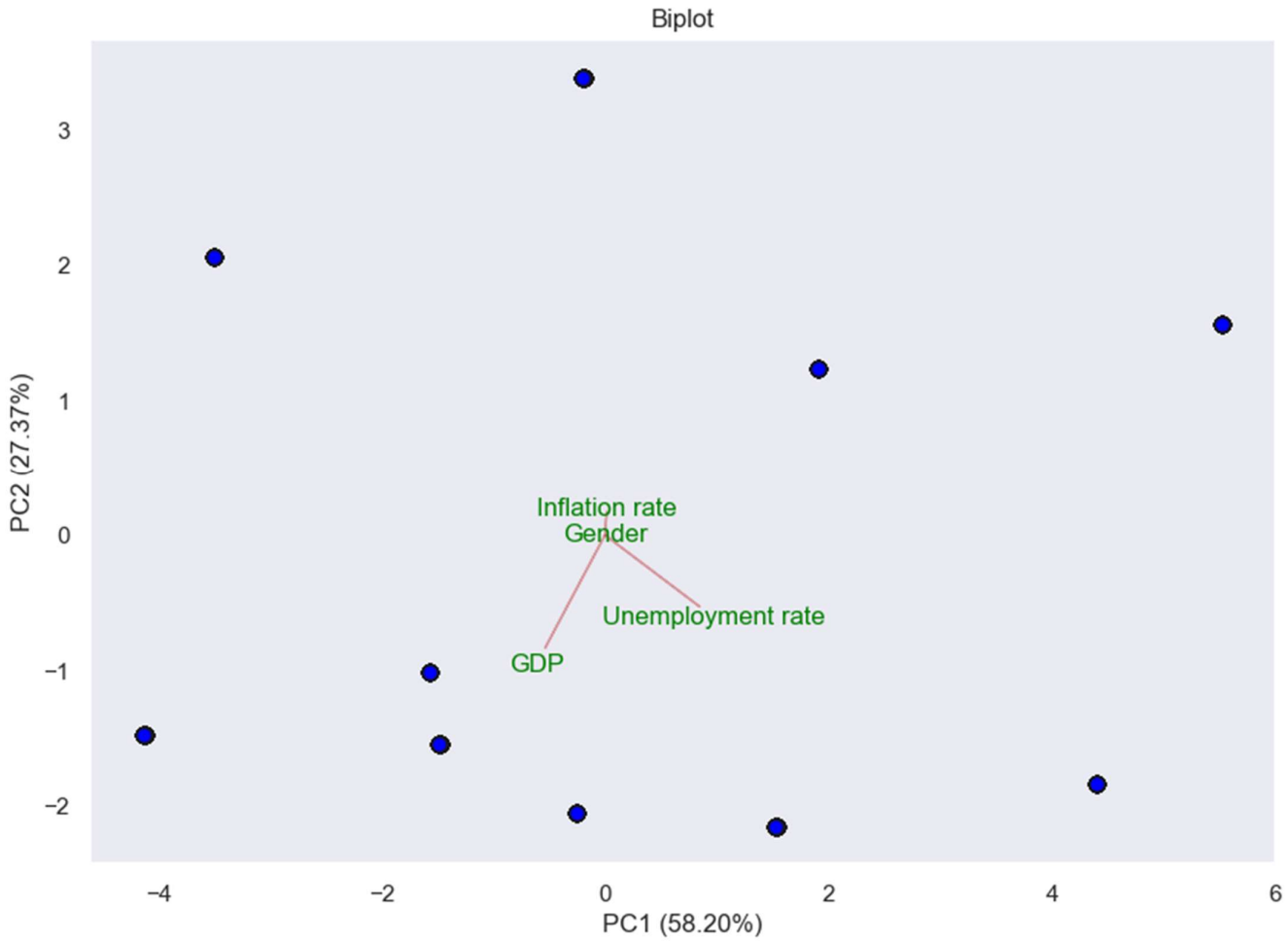
Summary:

- There is a general trend where the mean drop status increases with age. This could indicate that older individuals are more likely to experience "drop" events compared to younger individuals.
- While there are some variations, the overall trend of increasing drop status with age seems to hold true for both genders. However, there might be slight differences in the magnitude of the

increase between genders. For the 'Female' (0) gender, the dropout rate decreases as the age group increases from 'advanced' to 'senior'. This trend is not there for the male gender.

The 'adult' age group i.e ages (25-35) have the highest dropout rate for both genders

Biplot for Principal Component Analysis Visualization



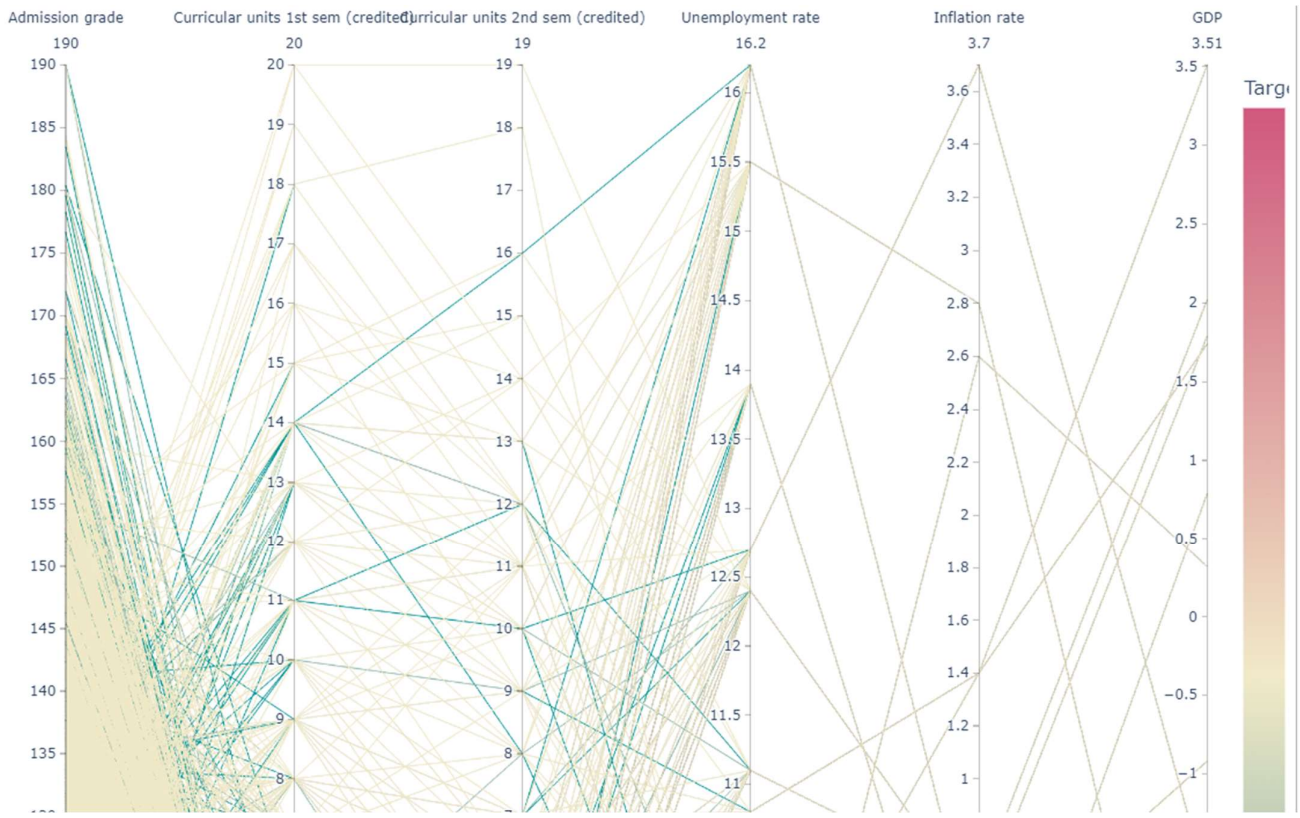
Summary:

- A few number of variables were selected for the biplot because too much features will lead to congestion of the plot. The features selected are the 'Unemployment rate', 'Inflation rate', 'GDP' and the 'Gender' column.
- These points are plotted according to the values of the first and second principal components (PC1 and PC2).

- PC1 explains 56.01% of the variance in the data, while PC2 explains 26.36%. Together, they capture around 82% of the variance, making this a good representation of the data.
- The red arrows represent the loading vectors, which show how the original variables contribute to the principal components.
- The direction and length of the arrows indicate the correlation between the variables and the principal components:
    - Longer arrows indicate variables that contribute more to that principal component.
    - Shorter arrows indicate variables that contribute less.
- In this case, the variables are labeled as 'Inflation rate', 'Application order', 'GDP', 'Unemployment rate', etc.
- Application order, Inflation rate, and Unemployment rate contribute more to Principal component 1 (along the x-axis) because their vectors are more aligned with this component.
- GDP contributes to both PC1 and PC2 but is more aligned with PC2 (y-axis).
- Variables like Application order and Inflation rate are clustered together, which indicates that these variables may be positively correlated with each other.
- GDP has a lower influence compared to other variables based on its shorter arrow.

Parallel Correlation Plot

A parallel correlation plot was built using four variables, 'Admission grade', 'Curricular units 1st sem (credited)', 'Curricular units 2nd sem (credited)', 'Unemployment rate', 'Inflation rate', 'GDP' to see how they vary with the target column



Summary:

This parallel coordinates plot provides a rich view into the relationships between student performance and various academic and economic factors. It appears that students with higher admission grades tend to follow a more predictable and stable pattern across the board. Their performance is more consistent, with many of them successfully completing a higher number of curricular units in both the first and second semesters. This could suggest that students who enter with stronger academic credentials are more likely to stay on track academically.

On the other hand, we see that when the unemployment rate is high, students' paths tend to scatter more. This might imply that during tough economic times, students face more challenges, possibly due to financial pressures, which could lead to more varied academic outcomes, including the likelihood of dropping out. Similarly, while inflation doesn't have as direct of an impact, periods of higher inflation also show a broader spread of paths, hinting at economic stress potentially affecting students' ability to perform or remain in school.

As we move across the variables, the GDP seems to represent a more stable influence. In times of higher GDP, students' performance appears to stabilize, whereas lower GDP seems to correlate with a broader range of outcomes. This could indicate that when the economy is doing well, students are better able to focus on their studies without external financial worries.

Finally, looking at the overall picture, there's a clear pattern that students who perform better or are less likely to drop out follow more consistent paths, particularly in terms of academic performance and external economic conditions. Conversely, those at risk of dropping out show a lot more variation in how they navigate through both academic and economic hurdles

New Hypotheses Generated

Based on the findings, several new hypotheses were generated for future investigation:

- Can parent's occupation affect dropout rates?
- Can parent's qualification affect dropout rates?