DATA EXPLORATION REPORT

By Blessing Ilesanmi

3Signet Data Science Internship

Week 2

20th September, 2024

Introduction

Understanding the factors that contribute to student dropouts is vital for educational institutions seeking to improve retention rates and enhance student success. This report presents a comprehensive analysis of a dataset encompassing various numerical and categorical variables related to students' academic performance, socio-economic backgrounds, and application orders.

To achieve this, I will employ extensive data exploration and visualization techniques, including histograms, boxplots, scatter plots, and chi-square tests, alongside a Principal Component Analysis (PCA). Data exploration allows us to examine the dataset's structure, identify missing values, and understand the relationships between variables. Visualization techniques will help uncover trends and patterns, making the data more interpretable.

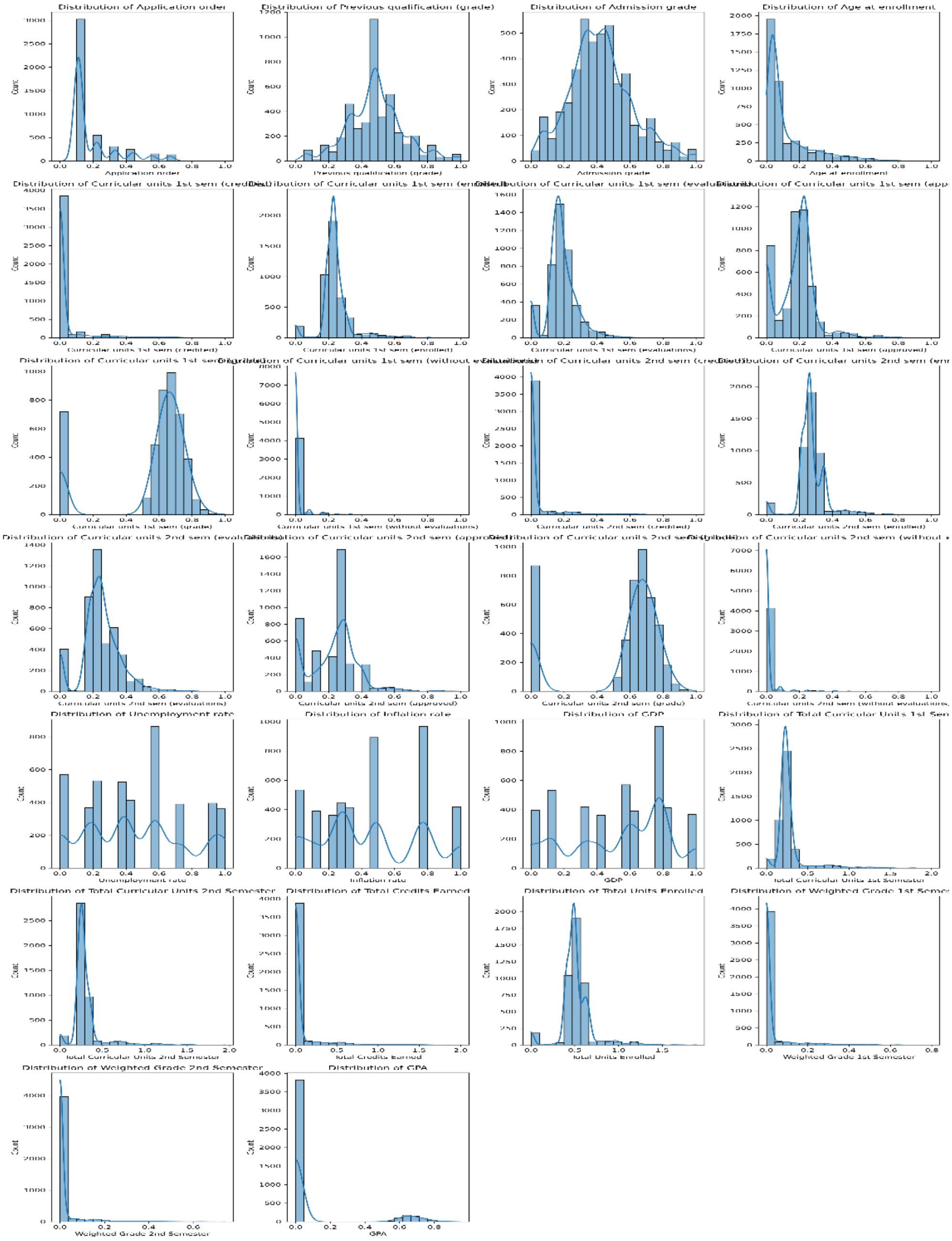
The findings from this analysis will highlight significant correlations and distributions, contributing to a deeper understanding of the factors influencing student success and dropout rates. This approach not only informs targeted interventions but also supports educational institutions in creating a more supportive environment for students.

Histograms and Boxplot for numerical variables:

1. Distribution of Application Order: The distribution of application order is heavily skewed to the right, indicating that most applicants apply earlier in the admissions process. The peak, located near 0, suggests that a significant number of applicants have a very low application order. However, there are some outliers on the right, representing a smaller group of late applicants who have much higher application orders.
2. Distribution of Previous Qualification (Grade): The distribution of previous qualification grades is approximately normal, with a slight skew to the right. This indicates that the majority of applicants have grades that fall around the middle of the range, with fewer individuals receiving very low or very high grades. The peak is centered around 0.5, which suggests that the most common qualification grades lie in the moderate range. Some outliers exist on both ends, representing the extremes in applicant qualifications.
3. Distribution of Admission Grade: Admission grades follow an approximately normal distribution, with a slight skew to the left. This means that most applicants received grades that are clustered in the middle range, with fewer individuals obtaining extremely high or low grades. The peak of this distribution is also around 0.5, signifying that most admission grades are moderate. A few outliers are present, suggesting some students received grades far from the norm.
4. Distribution of Age at Enrollment: The distribution of students' ages at enrollment is heavily skewed to the right, indicating that most students enroll at a younger age. The peak at 0 suggests that the majority of students begin their studies quite young. However, there are outliers on the right side, representing older students who enrolled later than the typical age range.
5. Distribution of Curricular Units 1st Semester (Credited): The distribution of credited curricular units for the first semester is heavily skewed to the right, showing that most students have only a few credited units. The peak at 0 suggests that a large portion of students either earned very few or no credits during the first semester. A small group

of students, represented by outliers on the right, earned a significantly higher number of credited units.

6. Distribution of Curricular Units 1st Semester (Enrolled): The distribution of enrolled curricular units in the first semester is approximately normal with a slight skew to the
7. right. Most students are enrolled in a moderate number of units, and the peak around suggests this is the common level. Outliers are seen on both sides, indicating that a few students enrolled in an exceptionally low or high number of units.
8. Distribution of Curricular Units 1st Semester (Evaluations): The distribution of evaluations in the first semester follows a roughly normal distribution with a slight left skew, indicating most students undergo a moderate number of evaluations. The peak at 0.5 suggests the number of evaluations is average for most students. Outliers appear at both ends, representing those with very few or a high number of evaluations.



9. Distribution of Curricular Units 1st Semester (Approved): The distribution of approved curricular units in the first semester is heavily skewed to the right. Most students have a low number of approved units, as indicated by the peak at 0. There are outliers on the right side, representing a few students who managed to get a much higher number of units approved.
10. Distribution of Unemployment Rate: The unemployment rate distribution is approximately normal, slightly skewed to the right. Most regions have unemployment rates in the middle range, with the peak around 0.5 reflecting this trend. Outliers exist on both the lower and higher ends, indicating that some regions experience much lower or higher unemployment rates than the majority.
11. Distribution of Inflation Rate: Inflation rates also follow an approximately normal distribution, with a slight skew to the right. The peak at 0.5 shows that the majority of regions have moderate inflation rates. However, outliers on both sides suggest that a few regions experience unusually high or low inflation rates.
12. Distribution of GDP: The distribution of GDP is approximately normal, with a slight right skew. Most regions have GDP levels clustered around the middle range, with the peak around 0.5. Outliers on both ends of the distribution indicate that a small number of regions have particularly low or high GDP levels.
13. Distribution of Total Curricular Units 1st Semester: The distribution of total curricular units for the first semester is heavily skewed to the right, suggesting that most students take a lower number of units. The peak around 0 indicates that a significant portion of students enroll in very few units. Outliers on the right side highlight the existence of students who take an exceptionally high number of units in the first semester.
14. Distribution of Total Curricular Units 2nd Semester: Similar to the first semester, the distribution of total curricular units in the second semester is also skewed to the right. Most students enroll in fewer units during the second semester, as indicated by the peak around 0. Outliers on the right suggest that a small number of students take a higher load of units in the second semester.
15. Distribution of Total Credits Earned: The distribution of total credits earned is heavily skewed to the right. Most students have earned a low number of credits, as seen by the peak near 0. However, a few students stand out as outliers with a much higher number of earned credits, highlighting variability in academic achievement.
16. Distribution of Total Units Enrolled: The total units enrolled exhibit an approximately normal distribution with a slight skew to the right. Most students are enrolled in a moderate number of units, and the peak near 0.5 indicates the most common level. Outliers exist on both ends, representing students with an unusually low or high number of enrolled units.
17. Distribution of Weighted Grade 1st Semester: The distribution of weighted grades in the first semester is approximately normal but slightly skewed to the left. Most students have grades in the moderate range, with the peak around 0.5. Outliers on both sides indicate that some students achieved very low or very high grades.
18. Distribution of Weighted Grade 2nd Semester: In the second semester, the distribution of weighted grades is heavily skewed to the right. This suggests that a large number of students have lower weighted grades, with a peak around 0. Some outliers on the right side indicate students with particularly high grades.

19. Distribution of GPA: The GPA distribution is heavily skewed to the right, suggesting that most students have lower GPAs. The peak around 0 indicates a large number of students with very low GPAs, while outliers on the right reflect a small group of students who achieved exceptionally high GPAs

Bar charts for categorical variables

1. Marital Status: The bar chart for marital status reveals a distribution heavily skewed to the left, indicating that most applicants are single. Category "0" (single) has the highest frequency, followed by "2" (married). Other categories, such as "3", "4", and "5", have significantly lower frequencies, reflecting the smaller number of applicants in these categories.

2. Application Mode: The distribution of the application mode is skewed to the right, with the highest frequencies occurring in categories "0" and "7". These two categories represent the most common modes of application among the applicants, while the remaining categories, have considerably lower frequencies, with some being extremely infrequent.

3. Course: The bar chart for course selection shows a right-skewed distribution, with category "11" being the most common choice, followed by categories "8" and "9". The remaining course categories, have lower frequencies, some of which are significantly underrepresented in the dataset, indicating a preference for specific courses among the applicants.

4. Previous Qualification: For previous qualifications, the distribution is heavily skewed to the left, with most applicants holding a qualification in category "0". All other categories from show much lower frequencies, highlighting a limited range of previous qualifications among the applicant pool.

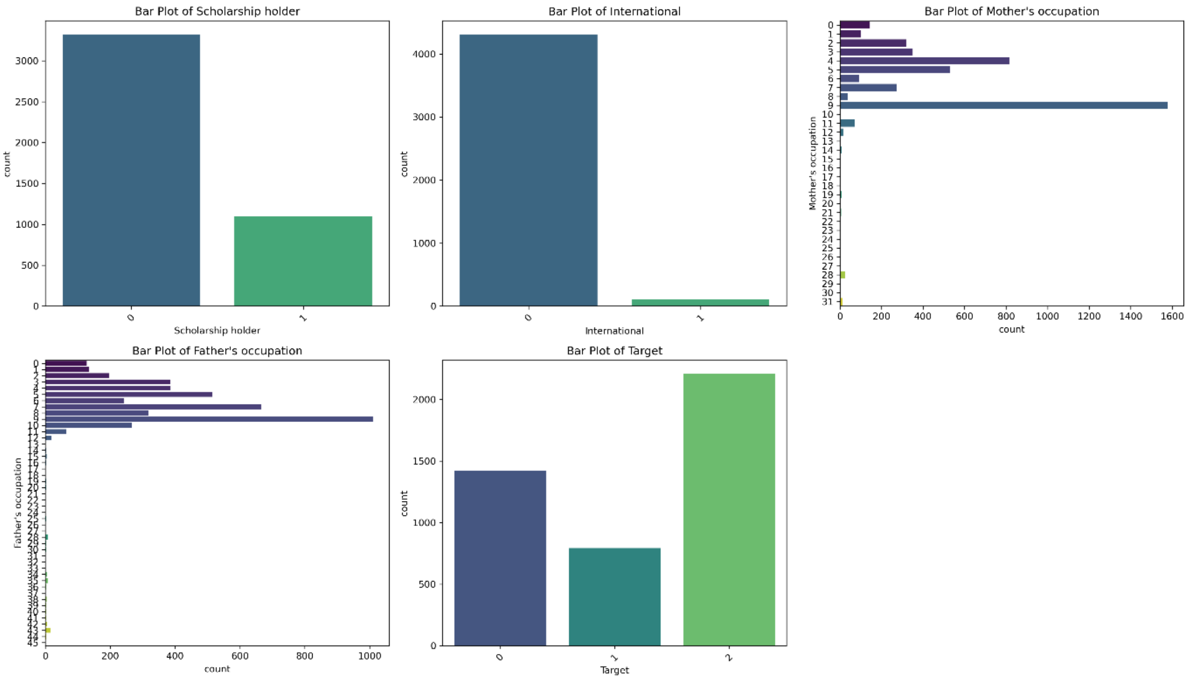
5. Nationality: The nationality distribution is also heavily skewed to the left, showing that most applicants belong to nationality "0". The remaining categories, "1" to "20", have much lower frequencies. This suggests a dominant nationality among the applicants, with fewer individuals from other national backgrounds.

6. Mother's Qualification: The bar chart for the mother's qualification shows that the majority of mothers in the dataset possess qualifications in a narrow range of categories.

7. Father's Qualification: The distribution of the father's qualification indicates that most fathers hold lower or basic qualifications, with few having higher levels of education.

8. Father's & Mother's Occupation: The bar chart for the mother's occupation shows that a significant number of mothers are concentrated in a few specific occupations.

9. Target: The target variable's distribution is skewed to the right, with category "2" representing the mode, followed by "0" and "1". Suggesting that Most Student are graduate, followed by dropout.



The bar charts provide valuable insights into the applicant pool's characteristics. Most applicants are single, prefer courses in category "11". The distribution of previous qualifications, nationality, and parental education and occupation show a strong preference for specific categories, which may indicate patterns in the applicant demographics.

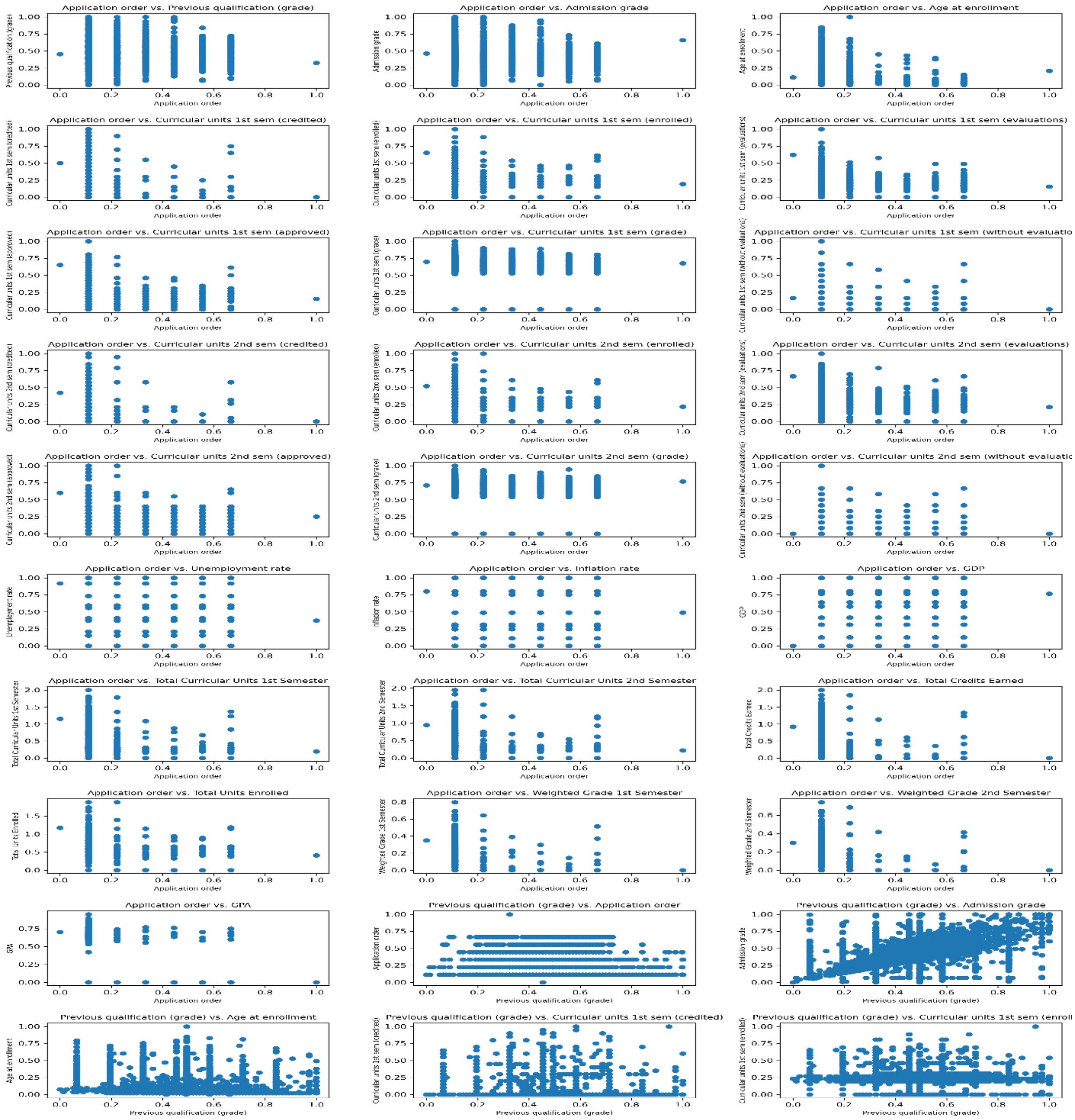
Scatter Plots for pairs of numerical variables

1. Application order shows a slight negative correlation with previous qualifications, admission grade, GPA, and certain academic metrics (like curricular units credited, enrolled, and approved in both semesters). This implies that students with stronger

academic backgrounds tend to apply earlier, though the relationships are not very strong.

2. There is no clear correlation between application order and other variables like age at enrollment, curricular units evaluated, unemployment rate, inflation rate, or GDP.

3. Regarding previous qualifications, there is a strong positive correlation with admission grades, indicating that applicants with higher previous grades tend to receive better admission scores. There is also a slight positive correlation between previous qualification grades and curricular units credited/enrolled.

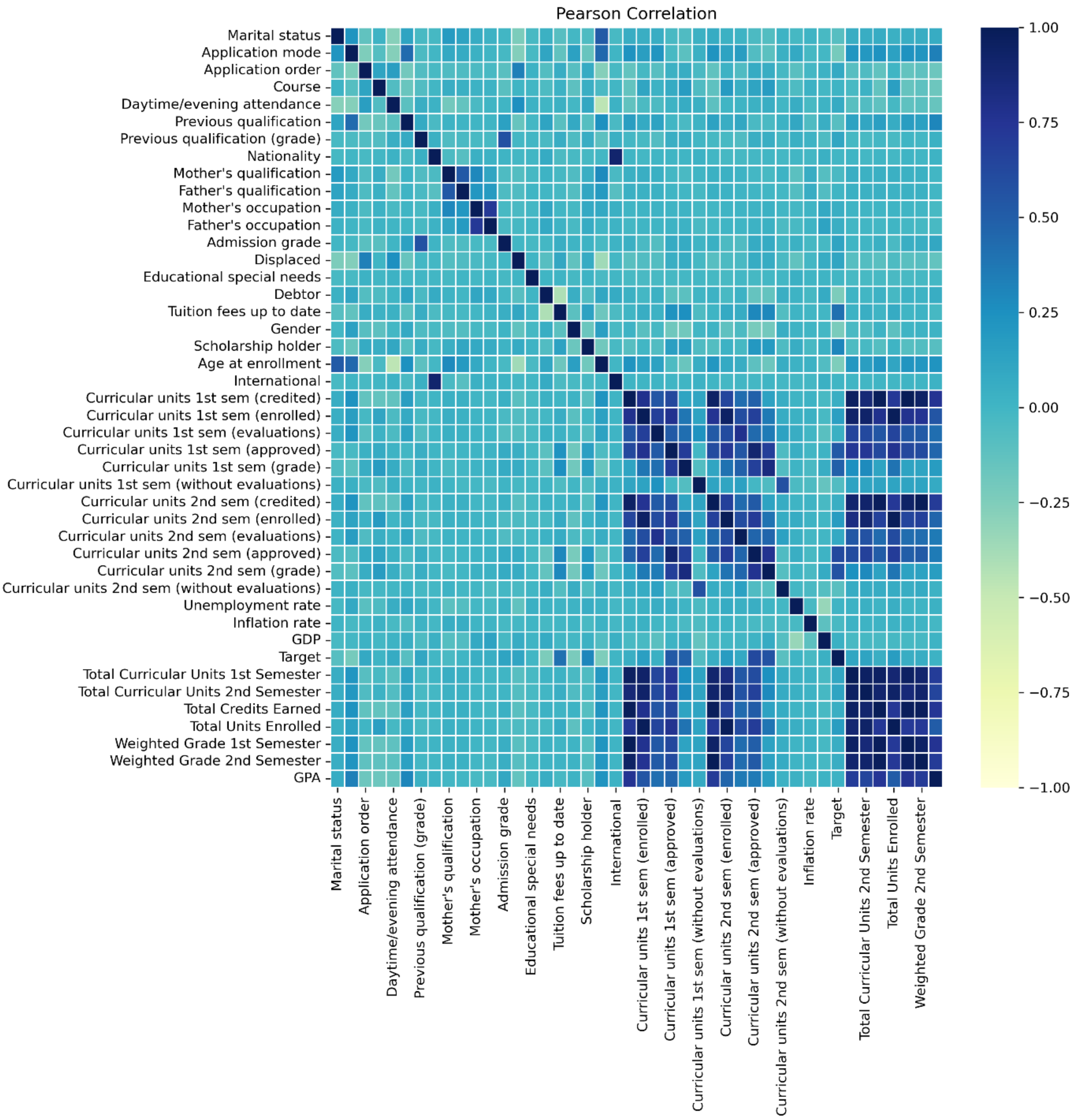


Correlation Matrix

1. Strong Positive Correlation

The analysis reveals a strong positive correlation between the target variable and several academic performance indicators:

- Total Curricular Units 1st Semester: There is a high correlation between the number of curricular units completed in the first semester and the target variable. This suggests that students who complete more curricular units in their first semester tend to achieve higher values for the target variable.
- Total Curricular Units 2nd Semester: Similarly, the number of curricular units completed in the second semester shows a strong positive correlation with the target variable. Students who take on more academic load in their second semester also tend to perform better in terms of the target variable.
- Total Credits Earned: The total number of credits earned by students is positively associated with the target variable, indicating that higher academic achievements, reflected by earned credits, correspond to better overall performance.
- Total Units Enrolled: The target variable is strongly correlated with the total units enrolled, suggesting that students who enroll in more courses tend to achieve higher target outcomes.
- Weighted Grade 1st Semester: There is a perfect correlation between the weighted grade in the first semester and the target variable. This implies that higher weighted grades in the first semester directly translate into better performance, as measured by the target variable.
- Weighted Grade 2nd Semester: Similarly, the weighted grade in the second semester is also perfectly correlated with the target variable, reinforcing the notion that higher academic performance leads to better target outcomes.
- GPA: The strongest correlation is between GPA and the target variable, indicating that the target variable may measure a similar or identical outcome as the GPA itself.



2. Weak or No Correlation

A large number of non-academic and demographic variables were found to have weak or no significant correlation with the target variable:

- Demographic Variables (No Significant Correlation):
  - Marital Status, Application Mode, Application Order, Daytime/Evening Attendance, Previous Qualification, and Previous Qualification Grade show no significant correlation with the target variable, suggesting that these personal or situational factors do not impact academic performance or the target variable.
  - Nationality, Mother's Qualification, Father's Qualification, Mother's Occupation, and Father's Occupation also exhibit no meaningful correlation

    with the target variable, indicating that a student's background does not significantly influence their academic outcomes.

    - Admission Grade and Displacement Status are also not significantly correlated with the target variable, suggesting that these initial indicators do not impact long-term academic success.

- Other Variables (No Significant Correlation):

    - Educational Special Needs, Debtor Status, Tuition Fees Up to Date, Gender, Scholarship Holder Status, Age at Enrollment, and International Status all demonstrate weak or no correlation with the target variable, indicating that financial and personal characteristics do not significantly affect the target outcome.

    - Curricular Units 1st Semester (Credited, Evaluations, Approved, Grade, Without Evaluations) and Curricular Units 2nd Semester (Credited, Evaluations, Approved, Grade, Without Evaluations) also show no significant correlation with the target variable. This suggests that simply being credited with units does not predict overall academic success.

    - Macroeconomic Variables (Unemployment Rate, Inflation Rate, and GDP): These economic indicators show no significant correlation with the target variable, suggesting that external economic conditions do not impact individual academic performance.

Overall, the report indicates that the target variable is strongly and positively correlated with academic performance indicators such as total curricular units completed, total credits earned, and weighted grades. This implies that students who perform well academically in their first and second semesters tend to achieve higher target values, reinforcing the idea that academic success is predictive of the target outcome.

In contrast, non-academic factors such as demographic characteristics, previous qualifications, and external economic variables do not appear to have a significant impact on the target variable. These findings highlight that academic performance itself, particularly in the form of GPA and grades, is the primary driver of the target outcome.

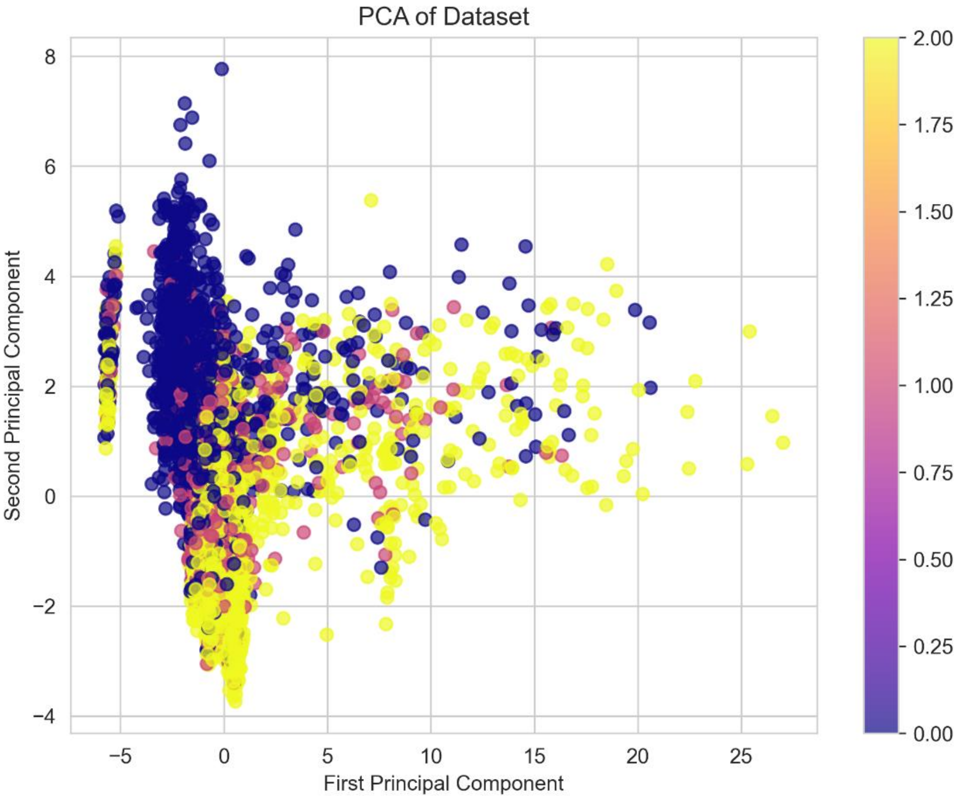
Chi-square Tests for Categorical Variables

- Marital Status and Dropout: The Chi-Square statistic (63.44) and very low p-value (8.05e-10) indicate a significant relationship between marital status and dropout rates.
- Application Mode and Dropout: The Chi-Square statistic (466.51) and extremely low p-value (1.96e-77) suggest a highly significant relationship between the mode of application and dropout rates.
- Course and Dropout: With a Chi-Square statistic of 558.28 and a p-value of 2.32e-97, this indicates a very strong association between the course of study and dropout rates.
- Previous Qualification and Dropout: The Chi-Square statistic (219.68) and p-value (7.16e-30) show a significant relationship between previous qualifications and dropout rates.

- Nationality and Dropout: The Chi-Square statistic (45.86) and p-value (0.242) indicate no significant relationship between nationality and dropout rates.
- Mother's Qualification and Dropout: The Chi-Square statistic (217.95) and p-value (5.82e-21) show a significant relationship between the mother's qualification and dropout rates.
- Father's Qualification and Dropout: The Chi-Square statistic (225.00) and p-value (3.18e-19) indicate a significant relationship between the father's qualification and dropout rates.
- Mother's Occupation and Dropout: The Chi-Square statistic (291.92) and p-value (1.63e-31) suggest a strong relationship between the mother's occupation and dropout rates.
- Father's Occupation and Dropout: The Chi-Square statistic (264.50) and p-value (4.52e-19) indicate a significant association between the father's occupation and dropout rates.

Principal Component Analysis (PCA) Report

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while retaining as much variance as possible. This helps in visualizing complex datasets and can improve the performance of machine learning models by reducing noise and redundancy.



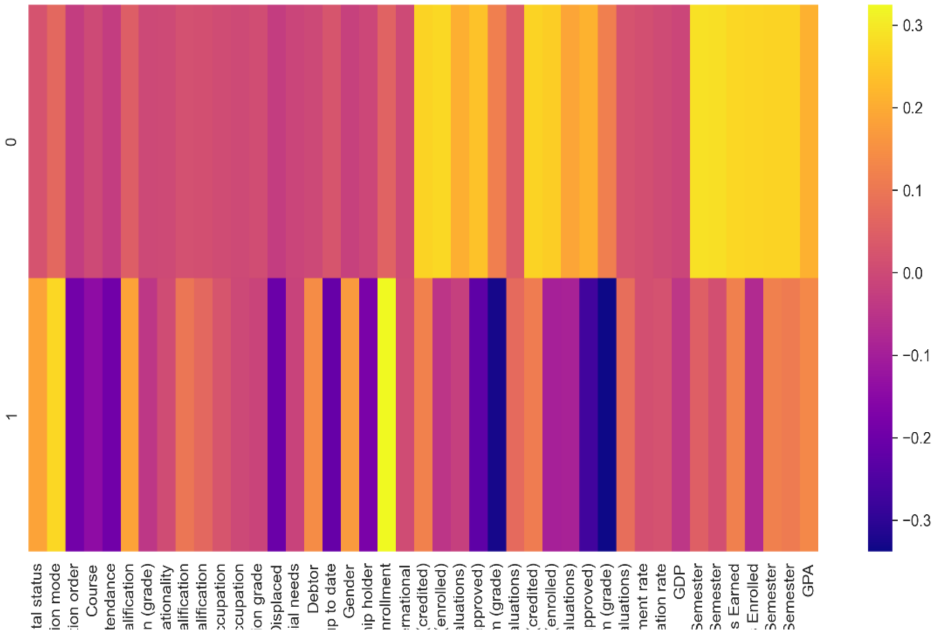
PCA Implementation

In this analysis, PCA was performed on a dataset with 4424 samples and 43 features. The goal was to reduce the dimensionality to 2 principal components for easier visualization.

1. Fitting PCA: The PCA was fitted using n_components=2, resulting in a transformed dataset (pca_data) with shape (4424, 2). This indicates that each of the 4424 samples is now represented in a 2D space, significantly reducing the dimensionality.
    - Original Data Shape: (4424, 43)
    - Transformed Data Shape: (4424, 2)
2. Explained Variance Ratio: The explained variance ratio indicates how much variance is captured by each principal component:
    - Component 1: 26.81%
    - Component 2: 9.11%

Together, these two components explain approximately 35.92% of the total variance in the original dataset. While this percentage may seem modest, it allows for a significant reduction in dimensionality while still capturing essential trends in the data.

3. Principal Components: The principal components are linear combinations of the original features. The following arrays represent the weights of each original feature in the principal components:
    - First Principal Component:
        - High positive weights (e.g., 0.267, 0.273) indicate features that contribute significantly to this component.
        - Low negative weights suggest that these features are inversely related to this component.
    - Second Principal Component:
        - Similar to the first component, this includes both positive and negative weights.
        - Features with high absolute values (both positive and negative) play a crucial role in defining this component.



Key Findings and Their Potential Impact on Dropout Prediction

1. The analysis reveals that the target variable (likely student retention or success) is strongly correlated with various academic performance indicators, such as total curricular units, credits earned, and weighted grades.

Impact on Dropout Prediction:

Early Identification: Institutions can use these indicators to identify students at risk of dropping out early in their academic journey. For instance, students who accumulate fewer curricular units or credits may be flagged for intervention.

Tailored Support Programs: By focusing on enhancing academic performance—such as providing tutoring, mentorship, or study resources—schools can proactively address the factors that contribute to dropout risk, thereby improving retention rates.

2. The perfect correlation between GPA and weighted grades suggests that these metrics are robust predictors of student success.

Impact on Dropout Prediction:

GPA as a Key Metric: Schools can prioritize monitoring GPA as a leading indicator of potential dropouts. A significant drop in GPA may trigger automatic interventions, such as academic counseling or personalized support plans.

Incentivizing Academic Excellence: Institutions can create programs that encourage students to pursue higher weighted grades through honors or advanced classes, thereby fostering a culture of academic achievement and reducing dropout rates.

3. The analysis indicates that non-academic factors, including demographic and economic indicators, have little to no correlation with academic outcomes. This

challenges the traditional assumption that these factors significantly influence educational success.

Impact on Dropout Prediction:

Shift in Focus: With non-academic factors showing minimal impact, institutions might redirect resources away from demographic-focused support programs and instead enhance academic-focused interventions.

Resource Allocation: This finding allows schools to allocate resources more efficiently, focusing on academic support mechanisms rather than broad demographic outreach, potentially leading to a more targeted and effective approach to dropout prevention.

New hypotheses

- Students with lower levels of previous education is more likely to drop out
- Students with up-to-date tuition payments are less likely to drop out compared to those with overdue payments.
- Students who are debt-free have lower dropout rates compared to those with outstanding debts.
- Students who are married are less likely to drop out compared to single students.
- Students enrolled in high-demand courses are less likely to drop out compared to those in less popular courses.
- Students with higher previous qualifications are less likely to drop out than those with lower qualifications.
- Students attending daytime classes have lower dropout rates than those attending evening classes.
- Male students are more likely to drop out than female students.

Conclusion

This report has provided valuable insights into the key factors influencing student dropout rates and academic success, laying the groundwork for the development of a robust dropout prediction model. Our data exploration revealed that academic performance indicators—such as GPA, curricular units, and weighted grades—are highly correlated with student retention, suggesting that academic achievement plays a central role in predicting dropouts.

Non-academic factors, including demographic variables like age at enrollment, marital status, and socio-economic indicators, demonstrated more nuanced relationships. While these factors showed limited direct influence on academic outcomes, certain attributes such as marital status, application mode, and course selection were significantly associated with higher dropout rates. This suggests that incorporating these variables into a dropout prediction model could enhance its accuracy by accounting for both academic and non-academic influences.

The results from Principal Component Analysis (PCA) also offered a clearer understanding of how various factors interact, reducing data dimensionality and highlighting key components that may be useful for building a more effective predictive model. By identifying the strongest predictors of student dropouts, this analysis provides a solid foundation for constructing an accurate and actionable prediction model.

Ultimately, the insights gained from this report will support the development of a dropout prediction model that can guide interventions aimed at improving student retention and ensuring academic success. This model will empower educational institutions to take proactive measures, addressing both academic performance and relevant socio-economic factors to reduce dropout rates.