Statistical Analysis Report

1. Descriptive Statistics Summary

Column name	count	mean	std	min	25%	50%	75%	max
Marital status	4424	1.18	0.61	1.00	1.00	1.00	1.00	6.00
Application mode	4424	18.67	17.48	1.00	1.00	17.00	39.00	57.00
Application order	4424	1.73	1.31	0.00	1.00	1.00	2.00	9.00
Course	4424	8856.64	2063.57	33.00	9085.00	9238.00	9556.00	9991.00
Daytime/evening attendance	4424	0.89	0.31	0.00	1.00	1.00	1.00	1.00
Previous qualification	4424	4.58	10.22	1.00	1.00	1.00	1.00	43.00
Previous qualification (grade)	4424	132.61	13.19	95.00	125.00	133.10	140.00	190.00
Nationality	4424	1.87	6.91	1.00	1.00	1.00	1.00	109.00
Mother's qualification	4424	19.56	15.60	1.00	2.00	19.00	37.00	44.00
Father's qualification	4424	22.28	15.34	1.00	3.00	19.00	37.00	44.00
Mother's occupation	4424	10.96	26.42	0.00	4.00	5.00	9.00	194.00
Father's occupation	4424	11.03	25.26	0.00	4.00	7.00	9.00	195.00
Admission grade	4424	126.98	14.48	95.00	117.90	126.10	134.80	190.00
Displaced	4424	0.55	0.50	0.00	0.00	1.00	1.00	1.00
Educational special needs	4424	0.01	0.11	0.00	0.00	0.00	0.00	1.00
Debtor	4424	0.11	0.32	0.00	0.00	0.00	0.00	1.00
Tuition fees up to date	4424	0.88	0.32	0.00	1.00	1.00	1.00	1.00
Gender	4424	0.35	0.48	0.00	0.00	0.00	1.00	1.00
Scholarship holder	4424	0.25	0.43	0.00	0.00	0.00	0.00	1.00
Age at enrollment	4424	23.27	7.59	17.00	19.00	20.00	25.00	70.00
International	4424	0.02	0.16	0.00	0.00	0.00	0.00	1.00
Curricular units 1st sem (credited)	4424	0.71	2.36	0.00	0.00	0.00	0.00	20.00
Curricular units 1st sem (enrolled)	4424	6.27	2.48	0.00	5.00	6.00	7.00	26.00
Curricular units 1st sem (evaluations)	4424	8.30	4.18	0.00	6.00	8.00	10.00	45.00
Curricular units 1st sem (approved)	4424	4.71	3.09	0.00	3.00	5.00	6.00	26.00
Curricular units 1st sem (grade)	4424	10.64	4.84	0.00	11.00	12.29	13.40	18.88
Curricular units 1st sem (without evaluations)	4424	0.14	0.69	0.00	0.00	0.00	0.00	12.00

Curricular units 2nd sem (credited)	4424	0.54	1.92	0.00	0.00	0.00	0.00	19.00
Curricular units 2nd sem (enrolled)	4424	6.23	2.20	0.00	5.00	6.00	7.00	23.00
Curricular units 2nd sem (evaluations)	4424	8.06	3.95	0.00	6.00	8.00	10.00	33.00
Curricular units 2nd sem (approved)	4424	4.44	3.01	0.00	2.00	5.00	6.00	20.00
Curricular units 2nd sem (grade)	4424	10.23	5.21	0.00	10.75	12.20	13.33	18.57
Curricular units 2nd sem (without evaluations)	4424	0.15	0.75	0.00	0.00	0.00	0.00	12.00
Unemployment rate	4424	11.57	2.66	7.60	9.40	11.10	13.90	16.20
Inflation rate	4424	1.23	1.38	-0.80	0.30	1.40	2.60	3.70
GDP	4424	0.00	2.27	-4.06	-1.70	0.32	1.79	3.51

The table provides a statistical summary for various features (columns) of the student's dataset.

1. **Count**: The total number of observations (rows) in the dataset for each feature. All features have 4,424 observations. This shows that there is no presence of missing values.
2. **Mean**: The average value of each feature.
3. **Standard Deviation (std)**: A measure of the spread or variability in the data.
4. **Min**: The minimum value for each feature.
5. **25%**: The 25th percentile value (first quartile), indicating that 25% of the data points are below this value.
6. **Median**: The median (50th percentile), showing the middle value of the data.
7. **75%**: The 75th percentile (third quartile), meaning 75% of the data points are below this value.
8. **Max**: The maximum value observed for each feature.

Insights:

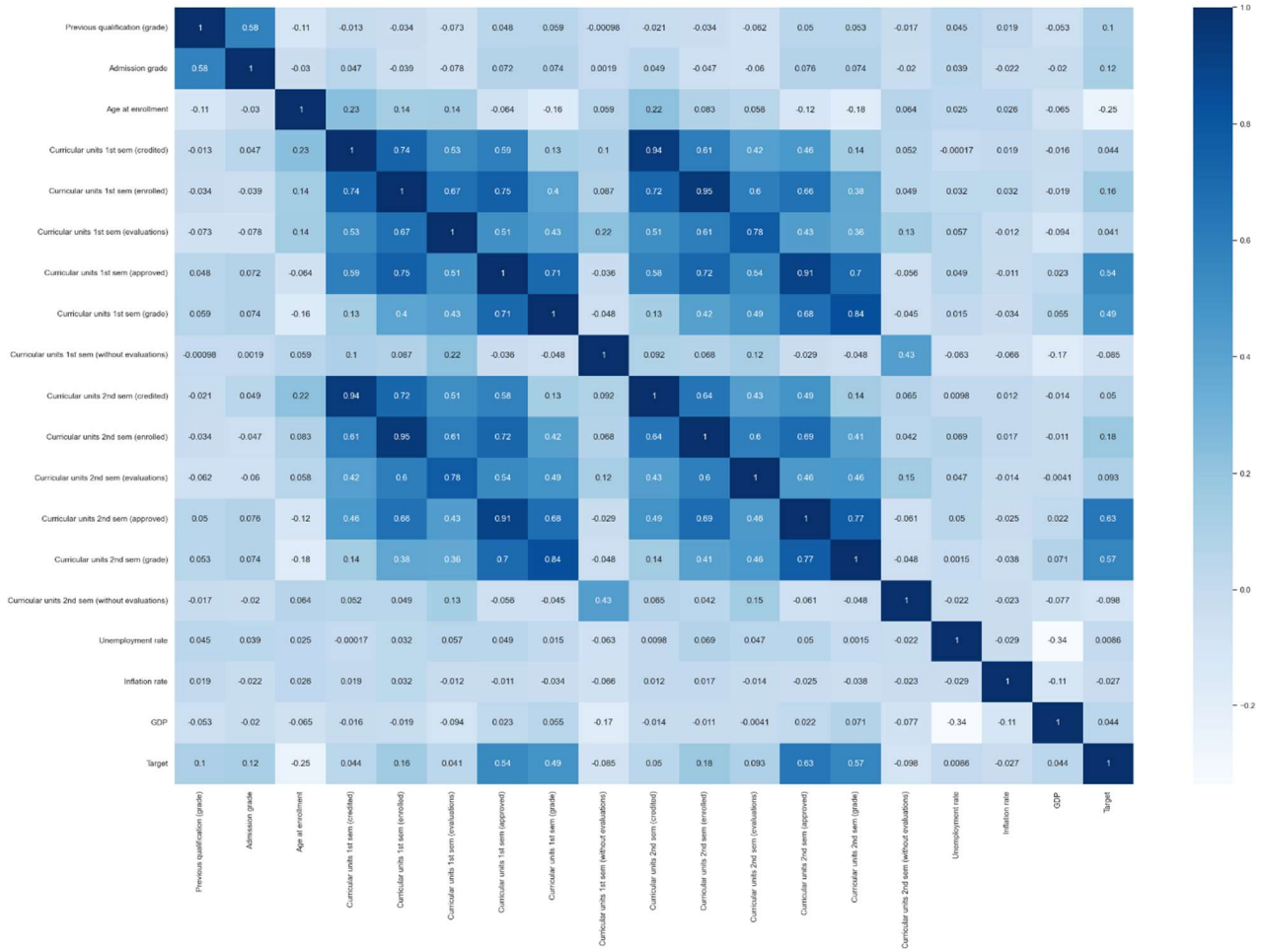
- **Marital status**: Most values seem to be concentrated on a specific category, as 75% of the data have a value of 1, while the max is 6.
- **Application mode**: It has a high standard deviation (17.48), with a mean of 18.67, indicating a wide range of modes with some large outliers (max = 57).
- **Application order**: Most applicants appear to be in the 1st or 2nd order, as the median (50%) is 1 and the third quartile is 2. The maximum value of 9 indicates that some applicants applied in later orders.
- **Daytime/evening attendance**: Most students attended during the day (median and 75th percentile = 1, with a max of 1).

- **Course**: The values range widely from 33 to 9991, with an average of around 8856. The high standard deviation (2063.57) suggests that different courses have widely varying codes or identifiers.
- **Admission grade**: The average admission grade is around 127 with a standard deviation of 14.48. The grades span from 95 to 190.
- **Age at enrollment**: The median age is 20 years, with a fairly large range of 17 to 70 years.
- **Curricular units (1st and 2nd semesters)**: This feature includes the number of credited, enrolled, evaluated, and approved curricular units, as well as the grades. The data shows a range of involvement in curricular units across students, with most students having a median grade of 12-13 and enrolling in 6-7 units per semester.
- **Unemployment rate**: The unemployment rate ranges from 7.6 to 16.2, with a median of 11.1%.
- **Inflation rate**: The inflation rate ranges from -0.8 to 3.7, with a median of 1.4%.
- **GDP**: The GDP values range from -4.06 to 3.51, with a median of 0.32, indicating some fluctuation in economic growth.

Conclusion:

- Many features, like **Marital status**, **Daytime/evening attendance**, **Debtor**, and **Tuition fees up to date**, have mostly binary data (0 or 1).
- Some features have significant variability, such as **Course**, **Application mode**, and **Mother's/Father's occupation**, suggesting a diverse range of values.
- **Curricular units** and **Admission grades** give insight into academic performance and engagement, while features like **Unemployment rate**, **Inflation rate**, and **GDP** provide economic context.

2. Correlation Heatmap:



Key Insights from the Heatmap:

- **Curricular units 1st sem (enrolled)** has a very high correlation with **Curricular units 1st sem (evaluations)** (0.95). This makes sense because the number of enrolled units directly influences the number of evaluations.
- **Curricular units 1st sem (evaluations)** has a strong positive correlation with **Curricular units 1st sem (approved)** (0.91). Students who evaluate units are likely to pass a portion of them.
- **Curricular units 2nd sem (enrolled)** and **Curricular units 2nd sem (evaluations)** (0.94) similarly show a very high correlation, indicating a consistent pattern of students enrolling and being evaluated in the second semester.
- **Curricular units 1st sem (approved)** is strongly correlated with **Curricular units 1st sem (grade)** (0.84). This is logical as better grades are generally associated with successfully passing more units.
- **Previous qualification (grade)** has a moderate positive correlation with **Admission grade** (0.58), suggesting that students with higher previous qualifications tend to have better admission grades.

- **Curricular units 2nd sem (approved)** correlates positively with **Curricular units 2nd sem (grade)** (0.77), meaning that students who pass more courses in the second semester tend to achieve higher grades.
- **Age at enrollment** has a moderate negative correlation with the **Target** variable (-0.25), indicating that older students might have a lower probability of achieving the target outcome (graduation).
- **Curricular units 1st sem (approved)** shows a moderately positive correlation with the **Target** variable (0.54), meaning that students who pass more units in the first semester are more likely to achieve the desired graduation rate.
- **Curricular units 1st sem (grade)** also correlates positively with the **Target** (0.49), indicating that higher grades in the first semester contribute to better graduation rate.
- **Admission grade** has a positive but weaker correlation with the **Target** (0.12), suggesting that students with higher admission grades are somewhat more likely to succeed.
- **Unemployment rate** shows a slight positive correlation with the **Target** (0.08), implying a weak relationship between unemployment and the target outcome, although the effect is minimal.
- **GDP** has a very weak but positive correlation with the **Target** (0.04), indicating that economic conditions may slightly influence student success.
- **Curricular units 1st sem (without evaluations)** has a negative correlation with the **Target** (-0.085). This suggests that students who have units without evaluations in the first semester may be less likely to meet the target.
- **Curricular units 2nd sem (without evaluations)** shows a stronger negative correlation with the **Target** (-0.098), which implies that failing to evaluate units in the second semester could reduce the likelihood of graduation.
- The **Inflation rate** has little to no correlation with other features, indicating that it doesn't play a significant role in student performance or any of the academic-related features.

Conclusion:

- Academic performance in the first and second semesters, as indicated by the number of evaluations, approvals, and grades, is highly correlated with achieving the likelihood of graduation.
- Economic variables like **GDP** and **Unemployment rate** have minimal impact on student outcomes.
- Students' performance in curricular units (especially in the first semester) is a strong predictor of whether they achieve the target, while age and lack of evaluations are negatively related to success.

3. Hypothesis Test Results:

The implementation of the hypothesis test involves the creation of a new variable called 'drop_stats'(dropout status) which indicates is a particular student dropped out or not. 1 represents Dropouts, 0 represents Non Dropouts irrespective of the fact that they are currently enrolled or graduates.
For chi-square contingency table tests, the Phi's coefficient which tell the degree and direction of the association between the two variables was computed.

Summary of the 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 status	Chi-Square Test	H_0: The variables are independent H_1: The variables are not independent	266.318	1.98E-56	Reject H0

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).