Data Preprocessing Report

Data Quality Assessment:

Criteria	Assessment	Recommendation
Completeness	The dataset appears to have no missing values.	Continue to maintain data entry procedures to ensure completeness.
Consistency	Data types for some features are not appropriate, and some values are not within reasonable ranges.	Implement data validation rules to ensure consistency and prevent the entry of invalid values.
Accuracy	There is a huge value of reasonable values that are outliers. A datasheet would be more helpful to deduce and understand what some features represent.	The datasheet containing the description of the features should be sent so one can know what values should be referred to as outliers.
Relevance	Majority of the features appear to be relevant in the prediction of the target variable.	Re-evaluate feature relevance if the dataset's use case changes. The datasheet will help in understanding the relevance of each feature.
Uniqueness	The dataset contains no duplicated records; each student sample is unique.	No action needed for uniqueness.

1. Description of Data Cleaning Steps

Handling Missing Data:

- No missing values was found in the analysis of the dataset.

Renaming of Columns:

- Certain features like the ‘nationality’ feature was renamed to ‘nationality’ and the 'Daytime/evening attendance \t' was renamed to 'Daytime/evening attendance'.

Duplicate Removal:

- A scan for duplicate rows was performed. No duplicate values were found.

Data Type Conversion:

- Categorical columns, such as 'Marital status', 'Application mode', 'Application order', 'Course', 'Daytime/evening attendance', 'Previous qualification', 'Nationality', 'Mother's qualification', 'Father's qualification', 'Mother's occupation', 'Father's occupation', 'Displaced', 'Educational special needs', 'Debtor', 'Tuition fees up to date', 'Gender', 'Scholarship holder', 'International', 'Target' were converted into category types. This reduced memory usage and helped proper handling during machine learning encoding.

Outlier Removal:

- The Z Score method with a threshold of 3.5 was applied to numeric columns such as 'Age at enrollment', 'Curricular units 1st sem (credited)', 'Curricular units 1st sem (enrolled)', 'Curricular units 1st sem (evaluations)', 'Curricular units 1st sem (approved)', 'Curricular units 1st sem (grade)', 'Curricular units 1st sem (without evaluations)', 'Curricular units 2nd sem (credited)', 'Curricular units 2nd sem (enrolled)', 'Curricular units 2nd sem (evaluations)', 'Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (grade)', 'Curricular units 2nd sem (without evaluations)', 'Unemployment rate', 'Inflation rate', 'GDP'to detect and remove outliers. There is a huge presence of outliers found in the dataset. This proved to be a huge problem because, on manual inspection, these values appear to be consistent and meaningful. These outliers was transformed using the natural log transformation.

Label Encoding:

- The target variable was encoded with 0 representing students who dropped out, 1 representing students who are enrolled and 2 representing students who have graduated.

2. Summary of Data Quality Issues Encountered and Resolved

- **Incorrect Data Types:** Some categorical features were incorrectly stored as integers, which hindered proper data analysis and model interpretation. This led to improper handling of these features during preprocessing.

Resolution:

These integer-encoded categorical features to the appropriate categorical data types. This optimized memory usage and ensured correct treatment during one-hot encoding, improving the overall quality and interpretability of the dataset.

- **Presence of Outliers**: During data analysis, outliers were identified as a key issue, with extreme values skewing several variables and increasing variance. This led to skewed distributions and unreliable model performance, as outliers disproportionately influenced results.

Resolution: To address this, we applied log transformation to affected variables. This technique reduced the impact of outliers by compressing the data range, stabilizing variance, and improving the symmetry of the distributions.

- **Difficulty in Understanding Variable Scale:** During the analysis, we encountered difficulty in understanding the scale of certain variables, as different features were measured on varying scales. E.g the 'GDP' variable is represented in float when it should be represented in integers or as a percentage.

3. Justification for Chosen Data Transformation Methods

Logarithmic Transformation:

Log transformation is a widely used data transformation method, particularly beneficial for improving the interpretability and performance of machine learning models. It is used for the following reasons:

1. **Reduces Skewness**: Log transformation is effective at reducing positive skewness in data by compressing the range of larger values more than smaller ones. This helps normalize distributions, which many machine learning algorithms assume.

2. **Stabilizes Variance**: For data with heteroscedasticity (where the variance increases with the magnitude of the variable), log transformation can stabilize variance. This leads to more robust statistical models and improved predictive accuracy.

3. **Improves Linear Relationships**: In many cases, a log-transformed variable can reveal or enhance linear relationships between variables, which is crucial for linear regression and other parametric models.

4. **Handles Exponential Growth**: Log transformation is useful when data represents exponential growth, such as population growth, sales trends, or financial returns. It linearizes exponential trends, making analysis and forecasting more manageable.

5. **Reduces Outlier Impact**: By compressing large values, log transformation minimizes the influence of outliers without entirely removing them. This prevents extreme values from distorting analysis.

6. **Prevents Data Loss**: Unlike techniques that remove or cap extreme values (such as outliers), log transformation retains the entire dataset while reducing the disproportionate influence of large values. This ensures no loss of valuable data, preserving information while making it more suitable for analysis.

Visualizations

i. Histogram

A histogram was plotted with the skewness value to visualize the skewness of the dataset and this influenced the scaling method that was applied

ii. Outlier Detection (Boxplot of Admission Grade)

A boxplot was used to detect and visualize outliers in the 'Admission grade' column. This revealed the presence of extreme values, which were subsequently removed using the IQR method.