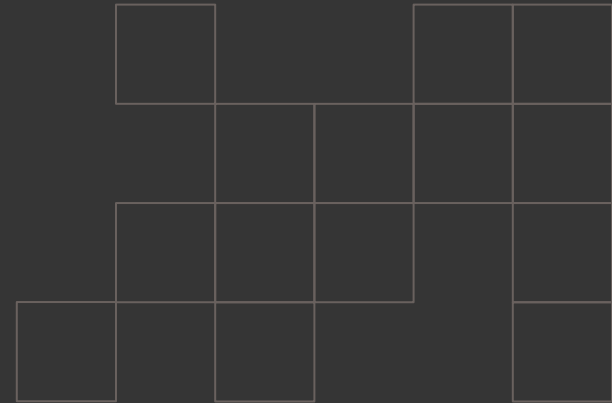


Predictive Analytics Midterm
10/14/2025
Report by: Rohan Mathew

Section 2:

Data Quality Assessment Report of Student Performance Factors Data Set



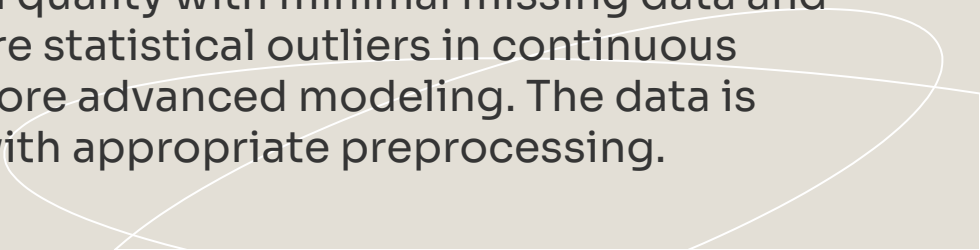


Executive Summary

This report systematically examines data quality issues in the Students Performance Factors dataset over 6000 student records with 20 variables related to academic performance factors.

The assessment reveals several data quality concerns that could impact analytical outcomes if not addressed.

The dataset demonstrates good overall quality with minimal missing data and high consistency. The main concerns are statistical outliers in continuous variables that should be addressed before advanced modeling. The data is suitable for most analytical purposes with appropriate preprocessing.



Data Cleaning



Missing Data

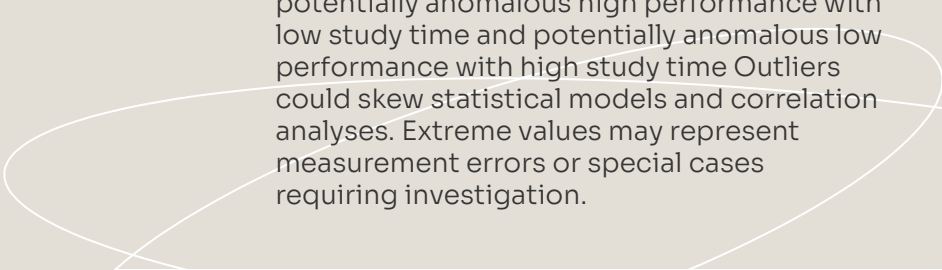
Minimal impact on analysis due to low prevalence, but could bias results if missingness correlates with unobserved factors.

Duplicates

Low risk of duplicate bias in analysis.

Outliers and Anomalies

potentially anomalous high performance with low study time and potentially anomalous low performance with high study time Outliers could skew statistical models and correlation analyses. Extreme values may represent measurement errors or special cases requiring investigation.



Data Cleaning Continued

Data Consistency

All categorical variables use consistent capitalization. No mixed formatting detected in text fields
High data consistency minimizes analysis errors. The physical activity scale anomaly should be verified.

Data Type and Validity Issues

Generally good data validity. Extreme values in Hours_Studied should be verified for accuracy.

Recommendations:

- 1. **Data Cleaning:**
 - Impute missing values using appropriate methods
 - Verify extreme values in Hours_Studied and Sleep_Hours
 - Standardize categorical variables if expanding dataset
- 2. **Analysis Considerations:**
 - Use robust statistical methods resistant to outliers
 - Consider stratified analysis for extreme cases
 - Validate findings across different data subsets
- 3. **Future Data Collection:**
 - Implement range validation during data entry
 - Add data quality checks for categorical variables
 - Consider logging data entry sources for traceability