

# Data Analysis Report

*Generated on 2025-04-23 23:26*

This report was automatically generated by AI Data Explorer, providing a comprehensive analysis of your dataset. It includes data quality assessment, key insights, statistical analysis, and actionable recommendations.

## Table of Contents

1. Dataset Overview
2. Executive Summary
3. Data Quality Assessment
4. Key Insights
5. Statistical Analysis
6. Advanced Insights
7. Recommendations

# 1. Dataset Overview

## General Information

Rows: 187  
Columns: 10  
Memory Usage: 0.04 MB

## Column Types

- int64: 7 columns
- object: 3 columns

## Data Preview

session	subject	max_total_mark	grade_a_star_th	grade_a_thresho
M_J	Add_Maths	160	150	120
O_N	Add_Maths	160	144	118
M_J	Add_Maths	160	140	113
O_N	Add_Maths	160	148	115
M_J	Add_Maths	160	141	119

*Note: This is just a preview. Some values may be truncated for display purposes.*

## 2. Executive Summary

### ### Comprehensive Summary of the Dataset

#### #### 1. Brief Overview of the Dataset

The dataset contains 187 rows and 10 columns, with information related to academic performance or grading thresholds across different subjects and sessions. The data includes various grade thresholds, maximum total marks, and contextual information such as the session and COVID-19 period.

#### #### 2. Key Insights about Structure, Completeness, and Quality

- **Structure**: The dataset is well-structured with a mix of categorical (``session``, ``subject``, ``covid_period``) and numerical variables (``max_total_mark``, various ``grade_*_threshold``).
- **Completeness**: There are no missing values in any of the columns, indicating that the dataset is complete.
- **Quality**: The data types are appropriate for the variables they represent. The presence of a good range of unique values in both categorical and numerical variables suggests a diverse dataset.

#### #### 3. Notable Patterns, Relationships, or Anomalies

- **Distribution of Categorical Variables**:
  - ``session`` has 3 unique values with "O\_N" and "M\_J" being the most frequent.
  - ``subject`` has 12 unique values, evenly distributed among the top 5 categories.
  - ``covid_period`` indicates that most data points are from "Pre\_COVID" and "Post\_COVID" periods.
- **Numerical Variables**:
  - ``max_total_mark`` and various ``grade_*_threshold`` show a decreasing mean and median as the grade threshold becomes lower (from A\* to E), which is expected.
  - The standard deviation decreases as the grade threshold lowers, indicating less variability in lower grade thresholds.
- **Correlations**:
  - Strong positive correlations exist among ``max_total_mark`` and all ``grade_*_threshold``, especially with ``grade_a_star_threshold`` (0.93) and ``grade_a_threshold`` (0.91).
  - The grade thresholds are highly intercorrelated, with correlations increasing as the thresholds get closer (e.g., ``grade_a_threshold`` and ``grade_b_threshold`` have a 0.98 correlation).
- **Anomalies**: No explicit anomalies are noted, but the relatively low number of observations for "July" session and "During\_COVID" period might limit analysis for these categories.

#### #### 4. Potential Use Cases or Analyses

- **Comparative Analysis**: Compare grading thresholds across different subjects or sessions to identify any biases or changes over time.
- **Impact of COVID-19**: Analyze the effect of the COVID-19 period on grading thresholds and maximum total marks.
- **Predictive Modeling**: Develop models to predict grade thresholds based on historical data, potentially including ``max_total_mark`` and ``subject`` as predictors.
- **Threshold Standardization**: Investigate the possibility of standardizing grade thresholds across subjects or sessions.

#### #### 5. Recommendations for Further Exploration

- **\*\*Segmented Analysis\*\***: Perform analyses segmented by `subject` or `covid\_period` to uncover more nuanced patterns.
- **\*\*Regression Analysis\*\***: Conduct regression analysis to understand the relationship between `max\_total\_mark` and various grade thresholds, potentially including interaction terms for `subject` or `covid\_period`.
- **\*\*Visualization\*\***: Utilize visualization techniques (e.g., scatter plots, bar charts) to better understand the distribution of variables and their relationships.
- **\*\*Outlier Detection\*\***: Although not immediately apparent, further statistical or visual inspection for outliers in numerical variables could be beneficial.

This dataset offers a rich source of information for understanding grading practices and their evolution, particularly in the context of the COVID-19 pandemic. Further exploration can provide insights into how grading thresholds are determined and how they might be standardized or predicted.

### 3. Data Quality Assessment

**Data Completeness: 0.00% missing**

No missing values detected in the dataset.

## 4. Key Insights

### Key Correlations

1. grade\_a\_star\_threshold and max\_total\_mark have a strong positive (0.93)
2. grade\_a\_star\_threshold and grade\_a\_threshold have a strong positive (0.99)
3. grade\_a\_star\_threshold and grade\_b\_threshold have a strong positive (0.95)
4. grade\_a\_star\_threshold and grade\_c\_threshold have a strong positive (0.73)
5. grade\_a\_threshold and max\_total\_mark have a strong positive (0.91)
6. grade\_a\_threshold and grade\_b\_threshold have a strong positive (0.98)
7. grade\_a\_threshold and grade\_c\_threshold have a strong positive (0.79)

# 5. Statistical Analysis

## Numerical Column Statistics

### Statistics for 'max\_total\_mark'

Measure	Value
Mean	141.42
Std Deviation	35.06
Minimum	30.00
25th Percentile	110.00
Median	150.00
75th Percentile	160.00
Maximum	200.00
Skewness	-0.40 (Approximately Symmetric)

### Statistics for 'grade\_a\_star\_threshold'

Measure	Value
Mean	111.29
Std Deviation	33.07
Minimum	25.00
25th Percentile	91.00
Median	110.00
75th Percentile	128.00
Maximum	188.00
Skewness	0.05 (Approximately Symmetric)

### Statistics for 'grade\_a\_threshold'

Measure	Value
Mean	93.65
Std Deviation	26.68
Minimum	21.00
25th Percentile	78.00
Median	93.00
75th Percentile	107.00
Maximum	160.00
Skewness	0.00 (Approximately Symmetric)

### Statistics for 'grade\_b\_threshold'

Measure	Value
---------	-------

Mean	74.61
Std Deviation	19.04
Minimum	17.00
25th Percentile	66.00
Median	74.00
75th Percentile	87.00
Maximum	117.00
Skewness	-0.41 (Approximately Symmetric)

Statistics for 'grade\_c\_threshold'

Measure	Value
Mean	56.17
Std Deviation	13.29
Minimum	13.00
25th Percentile	51.00
Median	58.00
75th Percentile	65.00
Maximum	86.00
Skewness	-0.87 (Left Skewed)

Note: Statistics for 2 more numerical columns are available in the interactive dashboard.

Categorical Column Analysis

Distribution of 'session'

Total unique values: 3

Category	Count
O_N	96
M_J	83
July	8

Distribution of 'subject'

Total unique values: 12

Category	Count
Biology	16
Business_Studies	16
Chemistry	16
Computer_Science	16
Economics	16



Distribution of 'covid\_period'

Total unique values: 3

Category	Count
Pre_COVID	95
Post_COVID	72
During_COVID	20

## 6. Advanced Insights

### High Correlation Between Grade Thresholds and Max Total Mark

The grade thresholds (A\*, A, B, C, D, E) are strongly correlated with the max total mark, indicating that the grading scale is generally proportional to the maximum possible score. The correlation is particularly strong for higher grades (A\*, A, B) with coefficients above 0.86.

*Confidence: high | Related columns: max\_total\_mark, grade\_a\_star\_threshold, grade\_a\_threshold, grade\_b\_threshold, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

### Grade Thresholds Become Less Correlated with Max Total Mark for Lower Grades

While higher grade thresholds (A\*, A, B) are strongly correlated with max total mark (coefficients: 0.93, 0.91, 0.86 respectively), the correlation weakens for lower grades (C, D, E) with coefficients of 0.65, 0.54, and 0.33 respectively. This suggests that the lower grade thresholds may be influenced by factors other than just the maximum total mark.

*Confidence: high | Related columns: max\_total\_mark, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

### Potential Inconsistency in Grading Scale Across Sessions or Subjects

The wide range of values for max total mark (30 to 200) and grade thresholds across different sessions and subjects suggests potential inconsistencies in the grading scale. This could be due to differences in exam difficulty, subject matter, or grading policies.

*Confidence: medium | Related columns: session, subject, max\_total\_mark, grade\_a\_star\_threshold, grade\_a\_threshold, grade\_b\_threshold, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

### Covid Period May Influence Grading Thresholds

The presence of a 'covid\_period' column suggests that the dataset may contain data from both pre-COVID and COVID-affected periods. Analyzing the impact of COVID on grading thresholds could reveal interesting insights into how the pandemic affected academic performance or grading policies.

*Confidence: medium | Related columns: covid\_period, grade\_a\_star\_threshold, grade\_a\_threshold, grade\_b\_threshold, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

### Data Quality Issue: Potential Outliers in Grade Thresholds

The minimum values for grade thresholds (e.g., grade\_a\_star\_threshold min: 25, grade\_e\_threshold min: 10) are significantly lower than their respective means and 25th percentiles. This could indicate outliers or data entry errors that warrant further investigation.

*Confidence: medium | Related columns: grade\_a\_star\_threshold, grade\_a\_threshold, grade\_b\_threshold, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

### Opportunity to Analyze Subject-Specific Grading Trends

The dataset contains data for multiple subjects. Analyzing grading thresholds and their relationship with max total mark across different subjects could reveal subject-specific trends or anomalies in grading practices.

*Confidence: high | Related columns: subject, max\_total\_mark, grade\_a\_star\_threshold, grade\_a\_threshold, grade\_b\_threshold, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

### Session-Specific Variations in Grading Thresholds

The dataset includes data across different sessions (e.g., M\_J, O\_N). Examining how grading thresholds vary across sessions could provide insights into seasonal or periodic patterns in academic performance or grading policies.

*Confidence: medium | Related columns: session, grade\_a\_star\_threshold, grade\_a\_threshold, grade\_b\_threshold, grade\_c\_threshold, grade\_d\_threshold, grade\_e\_threshold*

## 7. Recommendations

### Data Quality Recommendations

1. Your dataset is complete with no missing values - excellent data quality!
2. Investigate outliers identified in numerical columns, which may significantly affect statistical calculations.

### Analysis Recommendations

1. Explore the strong relationship between `grade_a_star_threshold` and `max_total_mark` further, as this may yield valuable business insights.
2. Consider dimensionality reduction techniques like PCA to identify underlying patterns in your numerical variables.
3. Perform group-based analysis to understand how key metrics vary across different categories.

### Next Steps

1. Use the interactive dashboard to explore visualizations that highlight key patterns in your data.
2. Leverage the chat interface to ask specific questions about your dataset.
3. Consider performing advanced statistical tests to validate key hypotheses.
4. Export insights and visualizations for presentation to stakeholders.