Comprehensive Student Performance Analysis

# Project Overview

This project delves into a dataset on student performance to understand the factors driving academic success and identify areas for improvement. By employing data cleaning, feature engineering, and advanced visualization techniques, the analysis aims to provide actionable insights for educators, students, and policymakers.

The project encompasses:

1. **Data Preparation:** Cleaning and optimizing data by handling missing values, detecting and managing outliers, and removing irrelevant columns.
2. **Feature Engineering:** Creating meaningful metrics such as performance scores, engagement categories, and attendance groups to enhance data interpretation.
3. **Visualization & Insights:** Using intuitive visualizations to explore key patterns, such as the impact of study hours, engagement, and attendance on performance.

# Python

## Data Cleaning

**Steps:**

1. **Handling Missing Values**:
   * Numeric columns (e.g., Study Hours per Week, Attendance Rate) were filled with the median.
   * Categorical columns (e.g., Parent Education Level, Library Card Number) were filled with the mode.
2. **Outlier Detection**:
   * Used the Interquartile Range (IQR) method to clip extreme values for numeric columns.
3. **Data Type Optimization**:
   * Converted categorical columns to the category data type.
4. **Dropped Unnecessary Columns**:
   * Removed columns like Contact Info and Library Card Number as they were irrelevant to the analysis.

## Feature Engineering

**New Features:**

1. **Performance Score**:
   * A weighted metric combining Previous Grades (50%), Attendance Rate (30%), and Study Hours per Week (20%).
2. **Engagement Category**:
   * Categorized Engagement Level into Low, Moderate, and High using defined thresholds.
3. **Attendance Group**:
   * Grouped students into Low, Moderate, and High attendance based on Attendance Rate.
4. **At-Risk Students**:
   * Identified students as "At Risk" if their Performance Score was below 50.

## Key Findings from Visualizations

**1. Academic Performance Distribution**

* Scores are relatively normally distributed, with a slight skew toward higher scores.

**2. Engagement Levels**

* Most students fall into the Moderate category, with fewer in Low or High.

**3. Attendance Groups**

* Majority of students are in the Moderate group, with fewer in Low or High.

**4. At-Risk Students**

* A smaller proportion of students are categorized as "At Risk".

**5. Correlation Matrix**

* Performance Score correlates moderately with Previous Grades, Attendance Rate, and Study Hours per Week.
* Engagement Level shows weak correlations with other numeric features.

## Additional Insights

**Average Performance Score by Engagement Category**

* Higher engagement levels are associated with higher performance scores.

**Study Hours by Attendance Group**

* Students in the High attendance group tend to have the highest average study hours.

## Final Steps

**Data Export:**

* Cleaned and processed data saved to cleaned\_student\_data.csv for further use.

**Recommendations:**

1. Focus interventions on students with Low engagement and attendance to improve overall performance.
2. Monitor study hours and attendance for students identified as "At Risk."
3. Explore deeper relationships between parent education levels and student performance.

# Observations about DataBase

The SQL schema reflects a well-structured relational database designed to support a detailed analysis of student performance. Key observations include:

A computer screen shot of a diagram

Description automatically generated

1. **Clear Data Organization**:
   * The schema normalizes data across multiple related tables, minimizing redundancy and improving data integrity.
   * Key entities like students, academicperformance, attendance, and engagement are logically separated, promoting modular and scalable design.
2. **Relationships and Integrity**:
   * Foreign key relationships ensure referential integrity between tables.
   * The use of a CountryID in the countries table allows for efficient handling of country-specific data.
3. **Comprehensive Feature Mapping**:
   * The academicperformance table captures detailed metrics like StudyHoursPerWeek, PerformanceScore, and relationships to engagement, attendance, and grades.
   * Engagement and attendance categories are effectively classified to allow granular analysis.
4. **Cleaned Data Integration**:
   * The cleaned\_student\_data table consolidates various features for direct analysis, integrating performance, engagement, and risk-related attributes.

## Challenges and Recommendations

1. **Missing Data Handling**:
   * Addressing missing data by filling numeric fields with the median and categorical fields with the mode reduces bias, but it assumes data is Missing Completely at Random (MCAR). Future analysis should explore imputation techniques sensitive to the dataset's characteristics.
2. **Outlier Handling**:
   * Clipping outliers using IQR is appropriate for initial cleaning, but domain knowledge could refine thresholds for specific variables like study hours and attendance rates.
3. **Categorization**:
   * Grouping variables like Engagement Level and Attendance Rate into categories helps simplify analysis but may lose detailed variability. Continuous variable analysis should be conducted in tandem.
4. **At-Risk Identification**:
   * The binary classification of "At Risk" students based on a performance score threshold is straightforward but may oversimplify. Multivariate analysis can improve identification accuracy.
5. **Normalization and Encoding**:
   * Normalizing numerical features and encoding categorical ones ensure compatibility with machine learning models. However, the pipeline could include feature selection to optimize processing.

## Strengths of the Approach

* The schema allows for flexible querying to address different analysis goals, such as understanding factors impacting engagement or attendance.
* Visualization of data provides actionable insights, such as targeting students with low engagement for interventions.
* Feature engineering enriches the dataset, making it suitable for advanced predictive modeling.

# Dashboard Insights

## Performance and Study Hours

* **Regional Performance**
  + Countries in Europe are color-coded based on performance scores, with darker shades representing higher scores.
  + Germany stands out with the highest performance score of 676,634, followed by France (574,734), the UK (455,896), and Spain (223,832).
* **Study Hours vs. Performance:**
  + The line chart depicts a correlation between weekly study hours and performance scores.
  + Peak performance is observed at approximately 19.5 hours per week, after which the performance declines despite increased study hours.

## Gender Analysis

* **At-Risk Students:**
  + Male students have a significantly higher count of "At-Risk" individuals compared to females, as shown in the bar chart.
  + However, a larger proportion of both genders are classified as "Not At-Risk."
* **Engagement Levels:**
  + The pie chart categorizes engagement into three levels: High, Moderate, and Low.
  + A majority of students (27,973) fall into the "Low" engagement category, indicating a potential area for improvement.
  + "High" engagement is the least common category, with only 3,153 students.
* **Study Hours by Gender:**
  + Male students contribute a total of 264,088 study hours, almost double that of female students (134,539).
  + This trend persists across all grade levels (10, 11, and 12).

## Key Observations

* **Study Hours and Performance Relationship:**
  + There is an optimal range of study hours (~19.5 hours per week) beyond which performance diminishes, suggesting diminishing returns or burnout from excessive study.
* **Gender Disparity:**
  + Male students dedicate more total study hours but are more frequently classified as "At Risk" compared to female students.
  + Females appear to achieve better performance outcomes with fewer study hours, indicating possible differences in study efficiency or external factors affecting male students.
* **Engagement Levels:**
  + A disproportionately high number of students are in the "Low" engagement category, which may negatively impact overall performance and risk status.

# Conclusion

The analysis provides a comprehensive overview of student performance, uncovering critical insights into the relationship between study hours, engagement levels, attendance, and gender-based disparities. It highlights areas that require targeted interventions, such as the large proportion of "Low" engagement students, the higher risk status of male students, and the diminishing returns from excessive study hours. By leveraging the findings, educators and policymakers can implement data-driven strategies to improve academic outcomes and create more equitable learning environments.