Comprehensive Report on Data Cleaning and Exploratory Data Analysis (EDA)

1. Introduction

This report summarizes the data cleaning and exploratory data analysis (EDA) conducted on the Students Grading Dataset using a Jupyter Notebook. The analysis includes preprocessing, outlier detection, univariate and bivariate analysis, and correlation studies to extract meaningful insights into student performance, study habits, and external factors influencing grades.

---

2. Data Cleaning and Preprocessing

The dataset was first examined for structural integrity, missing values, and inconsistencies. Key preprocessing steps included:

- **Handling Missing Values:**
    - **The dataset contained some missing values in specific categorical variables such as "Parent Education" and "Income Level."**
    - **Missing values were either imputed with the mode value or excluded based on their relevance to the analysis.**
- **Duplicate Removal:**
    - **Duplicate records were dropped, ensuring data integrity.**
- **Data Type Corrections:**
    - **Numeric and categorical variables were correctly formatted to facilitate analysis.**

---

3. Outlier Detection

- **Extreme outliers were identified in key variables such as exam scores, attendance, or participation.**
- **Outliers are handled through Z value method by finding IQR.**
- **The data distribution was relatively even across different metrics, ensuring that student performance metrics were well-balanced.**

---

4. Univariate Analysis (Numerical Variables)

4.1. Age Distribution

- **The age distribution is uniform (18-24 years old), suggesting a typical university-aged population.**

4.2. Attendance Percentage

- **Most students have an attendance rate of 70-80%, indicating moderate to high engagement.**

- **A small number of students have attendance below 60%, which might indicate frequent absences.**

4.3. Exam Scores (Midterm & Final)

- **Midterm and Final scores are distributed between 40 and 100.**
- **Scores are slightly right-skewed, meaning more students score higher rather than lower.**

4.4. Assignments, Quizzes, and Projects

- **Assignments & Quizzes: Most students score between 70-90.**
- **Project scores: Evenly spread, with most students scoring 60-100.**
- **Total Score: Follows a relatively normal distribution.**

4.5. Study Habits & Stress Levels

- **Most students study 10-30 hours per week, a common range in academic settings.**
- **Stress Levels: The majority experience moderate to high stress (4-8 levels on a scale of 10).**
- **Sleep Patterns:**
    - **Many students sleep 5-8 hours per night.**
    - **Students sleeping <6 hours may be at risk of burnout.**

---

5. Univariate Analysis (Categorical Variables)

5.1. Gender Distribution

- **The dataset is nearly balanced: 51% Male, 49% Female.**

5.2. Department Distribution

- **Computer Science (41.4%) has the highest representation, while Mathematics (10.1%) has the least.**

5.3. Grade Distribution

- **Grade A is the most common (29.9%), while Grade C is the least common (15.9%).**

5.4. Extracurricular Activities & Internet Access

- **Only 30.1% of students participate in extracurricular activities, indicating low engagement outside academics.**
- **89.7% of students have home internet access, ensuring minimal digital learning barriers.**

5.5. Parent Education & Family Income

- **38.4% of students have "Unknown" parental education data, which limits analysis.**
- **Most families are classified as Medium (39.5%) or Low income (39.7%), with only 20.9% in the high-income bracket.**

---

6. Bivariate Analysis (Relationships Between Variables)

6.1. Correlation Analysis

- **No strong correlations between numerical variables were found, suggesting that multiple factors contribute to performance.**
- **Study Hours vs. Total Score: Very weak correlation (~0.02), indicating that more study time does not guarantee better grades.**
- **Attendance vs. Total Score: Low correlation, meaning that simply attending classes does not ensure higher grades.**

6.2. Gender-Based Performance Trends

- **Participation Scores:**
  - **Female students tend to participate slightly more in class discussions.**
- **Stress Levels:**
  - **Female students report higher stress levels than males.**
- **Sleep Hours:**
  - **Males sleep slightly more on average.**

6.3. Department-Wise Performance

- **Computer Science students have the highest median attendance & total score.**
- **Civil Engineering students have the lowest median scores.**

6.4. Influence of Extracurricular Activities on Sleep

- **Students involved in extracurricular activities sleep slightly less than those who are not.**

6.5. Study Hours & Stress Levels Across Ages

- **Stress peaks at age 20, possibly due to academic pressure.**
- **Stress stabilizes after age 22, suggesting adaptation to workload.**

---

7. Multivariate Analysis

7.1. Heatmap Observations

- **Weak correlations across variables suggest no single dominant factor determines performance.**
- **Study Hours & Stress Levels are uncorrelated, meaning students with high stress do not necessarily study more.**

7.2. Pair Plot Analysis

- **Most variables exhibit a uniform spread, reinforcing weak direct relationships.**

7.3. Parent Education Level & Performance

- **Parental education does not strongly influence student grades.**

---

8. Conclusions & Recommendations

8.1. Academic Performance Trends

- **Most students perform well in midterms, finals, and projects, but quizzes show higher variability in scores.**
- **Attendance does not directly translate to better scores, suggesting self-study plays a role.**

8.2. Factors Affecting Student Success

- **Sleep and stress levels do not show strong relationships with academic performance, but students sleeping <6 hours should be monitored for burnout risks.**
- **Extracurricular activities have minimal impact on total score but may influence student well-being.**

8.3. Recommendations

- **Encourage low-attendance students to engage more.**
- **Improve quiz preparation strategies to reduce variability.**
- **Implement stress management programs for students with high stress (9-10 levels).**
- **Promote awareness of balanced sleep and study habits.**

Final Thoughts

This analysis reveals no single dominant factor determining academic success, highlighting the complexity of student performance. While grades are fairly well distributed, there are opportunities for improvement in quiz performance, engagement, and stress management strategies.